DAVID ANTONY SELBY

Statistical modelling of citation networks, research influence and journal prestige

DOCTOR OF PHILOSOPHY IN STATISTICS
DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK

JUNE 2020

Contents

List of Tables

List of Figures

Acknowledgements

I am extremely grateful to my supervisor, David Firth, for his constant guidance and kind support. I am indebted to Robin Ball, David Leadley, Peter Scott, Tim Bugg, Sascha Becker, Daniel Sgroi, Andrew Oswald, Robert MacKay, David Loeffler, Jonathan Forster, Jacob van Etten and others who offered valuable context on journal rankings in their respective disciplines. Thanks go to Thomson Reuters (now Clarivate Analytics) for providing Web of Science data in a convenient format, and Dave Santucci of Elsevier for his tips on using the Scopus database. Funding was provided by EPSRC grant `EP/M508184/1`.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. The work contained within is original, except as acknowledged, and has not been submitted previously for a degree at any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made.

Abstract

Standard approaches to measurement of the 'impact' of academic journals, or even sometimes of individual researchers or single research outputs, are typically not based on principled statistical methods for the analysis of citation data, through appropriate statistical models. Recent research has shown the value of such statistical modelling, for citations within a research discipline, for example in reproducing more faithfully the quality judgements of human assessors. In this project we study the strengths and weaknesses of statistical modelling approaches to citation-network data, and in so doing, uncover a deep theoretical connection between two otherwise unrelated journal ranking methods: PageRank and the Bradley–Terry model.

We extend the usual journal- or author-based metrics, by aggregating all publications in a given field into 'super-journals'. This permits modelling the exchange of citations between disciplines, raising the question: which scientific fields export the most intellectual influence, through recent research, to other fields? The relative merits of human and algorithmic field classifications are discussed. For this task, we propose a methodology of residual diagnosis for network community structures.

Finally, we investigate the extent to which the 2014 Research Excellence Framework's assessment of 'quality' of research outputs (rated 4*, 3*, 2* or 1*) was associated with the reputation of the journals in which those outputs were published. Submissions data are available, as are the aggregate scores for each university department, but the individual ratings for each paper are not. The research question is thus an 'ecological inference' problem attempting to estimate individual-level characteristics from aggregate data. Results are presented for several research fields.

To promote reproducibility and enable future research, the thesis includes a vignette on how to obtain citation network data from various databases, and is accompanied by R packages `scrooge` and `ref2014` to facilitate analysis.

Introduction

Citation analysis is a part of bibliometrics, the statistical analysis of written publications. It was originally intended for use by researchers and librarians to help make decisions about what to read and where to publish, given limited time and resources. However, the use of citation analysis has encroached into the realm of research assessment, with important implications for the careers of academics and allocation of research council funding, from institutions to the national level.

We start, in Chapter 1, by exploring the background of bibliometrics, and the debate surrounding its role in research assessment. A careful review reveals that these so-called statistical analyses of written work are perhaps not so statistical, with opaque and oft-manipulated data, questionable methodologies and little attempt at to quantify the uncertainty in calculations or natural variation in the statistics. In particular, the journal impact factor, flawed in principle, is abused in practice, which motivates a search for alternative ways of quantifying flow of influence via citations, if only to provide a least-worst option.

Two such alternatives, PageRank (eigenvector centrality) and the Bradley–Terry (quasi-symmetry) model, are competing ranking methods in bibliometrics. The Bradley–Terry model is a classical statistical method for ranking based on paired comparisons. The more recent PageRank algorithm ranks nodes according to their importance in a network.

Whereas Bradley–Terry scores are computed via maximum likelihood estimation, PageRanks are derived from the stationary distribution of a Markov chain. Recent work (Negahban et al., 2012; Maystre and Grossglauser, 2015) has shown maximum likelihood estimates for the Bradley–Terry model may be approximated from such a limiting distribution. However, this research overlooks fundamental work from Pinski and Narin (1976) in bibliometrics that provided the basis for PageRank.

In Chapter 2 we show—through relatively simple mathematics—a connection between paired comparisons and PageRank that exploits the quasi-symmetry property of the Bradley–Terry model, with direct implications for citation-based journal ranking metrics. We use the delta method to show that such an estimator is fully ef-

ficient when ranking similar-ability players in regular tournaments.

For a single research field, the work of Varin et al. (2016) demonstrated the value of statistical modelling of citations, for example in reproducing more faithfully the quality judgements of human assessors. Citation metrics are often criticised for their variation between disciplines, and some attempt to 'normalize' for this. However, the notion of what constitutes a 'discipline' is ill-defined; classifications based on human-curated lists of journals are often arbitrary, irreproducible and vulnerable to human error. Moreover, academic communities may evolve over time.

Community detection involves looking for groups in networks, and may be an algorithmic way of determining fields via citation data. Several different criteria are routinely used to measure success or stopping times for community detection algorithms. However, most of these have little statistical basis. But detecting lack of fit, outliers and unexplained structure is routinely done in generalised linear models by way of residual diagnostics. In Chapter 4, we propose using residual analysis of an implicit log-linear model to assess the quality of community detection results. This enables visualisation of uncaptured structure and, in citation analysis, the modelling of excess self-citation behaviour. These techniques are applied to a dataset of citations between statistical journals and their results discussed.

Given such a grouping of journals into a field, we can invoke the notion of 'super-journals' (Stigler, 1994) that represent multiple journals, merging the nodes of the citation network. Fitting a ranking model to these aggregated data can reveal the flow of influence between academic fields via citations. This has important implications for research assessments, which can often overlook the role of interdisciplinary impact. We find that over a short time window, citations tend to flow from pure subjects like mathematics towards more applied subjects such as medicine, rather than the other way round. By introducing a 'super-journal' to a within-field citation network, representing citation flow in and out of that field, we can also blur the distinction between siloed, local rankings and global rankings that ignore or normalize the behaviour of different fields, and see which journals that are prominent within their own community are perhaps less influential outside it, and vice versa. An interactive, animated visualisation of the intra- and inter-field rankings is presented to communicate the results.

A more pragmatic problem, considered in Chapter 5, is how to obtain citation data in the first place; it is no good producing interesting results that cannot be replication or useful methods that cannot be easily applied. Historically, bibliometric analyses relied on Web of Science data, but more recently alternatives have emerged, including Elsevier's Scopus database, Google Scholar and

Microsoft Academic, as well as the open-data initiative, the Open Citations Corpus. Several reviews have attempted to measure the 'coverage' of these various sources, but none provides a practical guide to constructing a citation network from each of them. We demonstrate, using simple reproducible R code (R Core Team, 2019) how to download and wrangle citation data from Scopus, Microsoft Academic and the Open Citations Corpus, show some example analyses and discuss the relative level of freedom each citation repository offers.

Finally, we look at research assessment in greater detail. The Research Excellence Framework (REF) is a periodic UK-wide assessment of the quality of published research in universities. The most recent REF was in 2014, and the next is currently scheduled to take place in 2021. The published results of REF2014 include a categorical 'quality profile' for each unit of assessment (typically a university department), reporting what percentage of the unit's REF-submitted research outputs were assessed as being at each of four quality levels (labelled 4*, 3*, 2* and 1*). Also in the public domain are the original submissions made to REF2014, which include—for each unit of assessment—publication details of the REF-submitted research outputs.

In Chapter 6 we address the question: to what extent can a REF quality profile for research outputs be attributed to the journals in which (most of) those outputs were published?

The data are the published submissions and results from REF2014. The main statistical challenge comes from the fact that REF quality profiles are available only at the aggregated level of whole units of assessment: the REF panel's assessment of each individual research output is not made public. Our research question is thus an 'ecological inference' problem, which demands special care in model formulation and methodology. The analysis is based, in the maximum likelihood case, on logit models in which journal-specific parameters are regularized via prior 'pseudo-data', and as an analogous Bayesian approach using Hamiltonian Monte Carlo and suitable regularizing priors. We develop a lack-of-fit measure for the extent to which REF scores appear to depend on publication venues rather than research quality or institution-level differences. Results are presented for several research fields.

Chapter 7 provides some concluding remarks.

1
Background

Standard approaches to 'objective' research assessment involve summarising counts of citations between academic journals. For various reasons, this might not be the right way to measure research quality or 'impact'. In this chapter, we briefly review a history of measures such as the journal impact factor and how the role they play in academia has changed. The information age has brought new opportunities for the way in which we conduct research, but many challenges as well; there are many vested interests and ulterior motives at work, which can make reliable measurement of researcher behaviour difficult.

It is a controversial subject, so any attempt to model research impact requires careful consideration of the data sources and methodology used, and potential for negative consequences.

1.1 Bibliometrics

There is only one way to know the quality of an academic paper: to read it thoroughly, equipped with a deep understanding of the subject area, or to speak to somebody else who has done so. But the vast scale and diversity of the academic literature makes this task impossible for librarians, who must decide which journals to purchase with limited funding, and for researchers, who must decide what to read and where to publish their work.

Every academic paper includes a bibliography to acknowledge the work of other authors. Citation analysis is the attempt to use these data to measure and compare the impact of published research. *Bibliometrics* is the statistical analysis of written publications more generally (Pritchard, 1969).

The use of bibliometrics to study the influence of research depends heavily on the assumption that academic journals are a primary means by which researchers communicate with each other and the wider world. Furthermore, citation analysis requires that references within an article be a reasonable record of the influences on a particular piece of work (MacRoberts and MacRoberts, 1996; Shema, 2013). Both of these points are long disputed; arguably there is more to the scientific method than contained within a published paper:

> 'There is a great deal in science that cannot conveniently, if at all, be included in publications… Not all, and hardly even the larger part, of scientific communication is carried on by published papers. To an extent much larger than is realized, the transference of scientific ideas from one set of scientific workers to another is effected by visits, personal contacts and letters.'
>
> Bernal (1939, pp. 119, 311)

Today such scholarly communication may also be accomplished via e-mail, the Web and social media; such informal collaborations are sometimes mentioned in acknowledgments but rarely in references (MacRoberts and MacRoberts, 2017). Moreover, data collected by third parties can be fundamental to research, yet databases are rarely cited (MacRoberts and MacRoberts, 2009). Meanwhile perverse incentives for adding citations are a well-documented phenomenon (Wilhite and Fong, 2012).

1.2 *The journal-publisher industrial complex*

Some authors do not criticise academic papers per se, but rather the system of collating them in scholarly journals that are usually closed-access. Schmitt (2014) called academic journals 'the most profitable obsolete technology in history'. This opinion is not a novel one, and indeed predates the electronic computer; Bernal (1939, p. 295) advocated abolishing journals[1] in favour of making 'the separate paper itself the unit of communication between scientists', arguing that the journal is 'obviously an inefficient way of distributing a large amount of scientific information', harking back to a time when it was possible to read every scientific publication there was—impossible by 1939, let alone today (see Figure 1.1).

Modern datasets are larger, no longer able to be tabulated within the pages of a physical journal. Organisations like DataCite have been established to encourage treatment of datasets as independent entities that are published and cited in the same way as journal articles (Brase, 2009).

Despite the falling cost of electronic distribution, large commercial publishers have increased, not decreased, their control of the academic literature over time (Larivière et al., 2015) and their profits (Schmitt, 2014), at odds with trends in most other media such as music and film. At the turn of the century, peer-to-peer file sharing suddenly made accessing music free and easy, changing expectations and exposing 'a willingness among many to use non-commercial means to obtain music if expectations were not met' (Bartsch, 2017). The music industry has since adapted, today offering user-friendly streaming services such as Spotify. Netflix, iPlayer and Amazon Prime represent similar responses in the film and TV industry.

Academic publishing has not yet had its 'Napster moment', though the infamous web site Sci-Hub[2] comes close: as well as opening up articles to users without subscriptions (especially

[1] Bernal (1939) proposed archiving articles on microform and printing paper copies on demand. At the time researchers were already exchanging article reprints directly with one another, a precursor of today's online pre-print servers like arXiv.

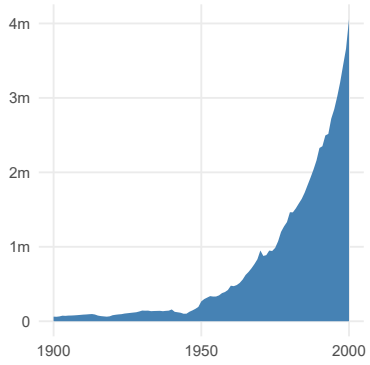


Figure 1.1: Number of academic articles published annually, 1900–2000, according to data from Microsoft Academic

[2] `https://www.sci-hub.tw`

in developing countries), Sci-Hub is widely used by researchers who already have access through their libraries, 'for convenience rather than necessity' (Bohannon, 2016). A recent study (Himmelstein et al., 2018) found that at least 85% of paywalled articles—as indexed by Crossref—are available on Sci-Hub, rising to 90% of Springer Nature, 95% of Wiley-Blackwell and 97% of Elsevier publications.

Elsevier has filed copyright infringement lawsuits against Sci-Hub creator Alexandra Elbakyan, but with little effect (Bohannon, 2016), and garnering little sympathy for publishers from the academic community. The site has been dubbed 'necessary, effective civil disobedience' (Brembs, 2016) whereas publishers call it 'illegal access… stealing content' (Cochran, 2016).

1.3 Open-access publishing

Currently the 'journal-affordability problem' limits access to research and therefore potentially limits the impact of that research. Shavell (2010) suggested, from a legal perspective, that abolishing academic copyright would in fact be socially desirable; his argument went as follows:

1. Academics do not benefit financially from copyright (they are paid little or no royalties) but benefit significantly in reputation and potential professional reward from publications.
2. Without copyright, publishers would have to impose publication fees to recover their costs.
3. If these costs were borne by authors, this would reduce incentives to publish, especially lower quality material (on the other hand authors could still post them on the internet).
4. If universities or funders subsidised publication fees instead (conditional on some threshold of quality for the venue of publication), this disincentive would disappear.
5. Research institutions have an incentive to subsidise fees because they a) want to encourage researchers to publish work, and b) would presumably save money on journal subscriptions and books.
6. Copyright-free research literature would increase access, reduce teaching costs and eliminate costs associated with protecting copyright.

In that same vein, support has grown for open-access publishing (Harnad et al., 2004). Open access publishing models include *gold*: publishing in an open-access journal that is free to read, but may charge authors fees for submitting articles (i.e. the model described by Shavell above); and *green*: where the journal itself is not open access, but allows authors to self-archive in an openly accessible repository. Here, 'free' sometimes—but not necessarily—means the article is published under a Creative Commons CC-BY licence and/or that the author retains copyright.

Nature Publishing Group and the Public Library of Science publish two open access 'mega journals', *Scientific Reports* and *PLoS One*, respectively, which have become the largest in the world (Davis, 2017). With an emphasis on scientific validity rather than importance, such journals 'could mop up the vast majority of published papers in the sciences', according to one *PLoS One* editor, who said: 'I think this is the death knell for the majority of "middling" journals and the large number of low-volume, low-profit, low-prestige journals' (Jump, 2011).

Some fields, such as machine learning research, are already so accustomed to open access publishing that attempts to introduce new paid-access journals, such as *Nature Machine Intelligence*, have met hostility (Coldewey, 2018), whilst the governments of France and Germany are at an impasse with Springer and Elsevier over subscription costs (Matthews, 2018).

The push for an open access model has also been exploited by unscrupulous publishers, who, driven by the lucrative incentive of article processing fees, accept submissions with little to no proper peer review (Beall, 2012; Bohannon, 2013). Librarian Jeffrey Beall maintained a controversial list of such 'predatory' journals and publishers. It was eventually shut down (Beall, 2017; Silver, 2017) but has been preserved by several anonymous contributors[3] (Chawla, 2018). Predatory publishing is not a problem with open access itself, however, rather with author fees. Indeed, many open access journals do not charge authors to submit articles—instead they are funded directly by research institutions (Rice, 2013).

[3] See for example `https://predatoryjournals.com`

But even for ostensibly nonpredatory publishers, the integrity of the gold open access model has been questioned. 'Because they aim to generate profits for their owners, gold (author-pays) open-access journals have a strong conflict-of-interest when it comes to peer review,' according to Beall (2017). 'They always want to earn money, and rejecting a paper means rejecting revenue.' One editor of *Scientific Reports* resigned after Nature Publishing Group introduced a fast-track privatised peer review service for authors willing to pay an extra fee (Bohannon, 2015).

A celebrated 'Subversive Proposal' (Harnad, 1994) called on authors to self-archive articles freely online. A review twenty years later (Poynder, 2014) suggested that full open access is 'vastly overdue', and that by imposing embargoes and offering 'hybrid gold' open access journals, publishers are trying to frustrate the growth of (green) open access in favour of '(Fool's) Gold OA', to sustain revenue.

In one study on green open access, Klein et al. (2018) compared published journal articles with their pre-prints, and found their content almost unchanged. Not all pre-prints in public repositories have been peer-reviewed, but as one researcher put it: "The primary role of traditional journals is to provide peer review and for that you don't need a physical journal—you just need an editorial board and an editorial process" (Harvard academic Sam

Gerschman, quoted by Schmitt, 2014).

Some journals, called *diamond* or *platinum* open access, are free to both readers and authors (Normand, 2018), sometimes providing little more than an 'overlay' for an existing public repository, such as arXiv (Moody, 2013). Organisations like Scholastica[4] and Épisciences[5] provide a minimalistic software for peer review of such articles. Publications are free to read and publication costs, if any are sometimes subsidised by institutional grants. Acceptance and peer review effectively gives a seal of approval to an article already published in an open repository.

[4] https://scholasticahq.com

[5] https://www.episciences.org

Though the peer review process dates back to the 18th century (Benos et al., 2007), today there is a problem of there being too many articles and too few people to review them all (Sipido et al., 2017). Crowd-sourced, post-publication peer review has been suggested as a solution (Harnad, 2014); for reviews of the topic see da Silva and Dobránszki (2014) and Knoepfler (2015).

Embracing this approach is the *WikiJournal of Medicine*[6], the first 'Wikipedia-integrated academic journal', created in 2014. Hosted by the Wikimedia Foundation, it has 'minimal requirements for technical maintenance by journal participants', no paywall for readers and no article submission fees for authors. Anyone can suggest edits to articles, from formatting and subediting to post-publication peer review (Shafee et al., 2017). Since 2016, the *WikiJournal of Science*[7] has played a similar role for articles in science, engineering and mathematics.

[6] https://en.wikiversity.org/wiki/WikiJournal_of_Medicine

[7] https://en.wikiversity.org/wiki/WikiJournal_of_Science

1.4 Citation metrics

One popular citation metric is the *impact factor*: the average number of citations a journal has received per published article over two years (Garfield, 1972).

Though widely used, the journal impact factor (JIF) is controversial (Seglen, 1997; Amin and Mabe, 2003; Garfield, 2006; Arnold and Fowler, 2011). It faces a number of criticisms, from ethical concerns over its effect on behaviour (van Wesel, 2015) to the possible undesirability of using the arithmetic mean to describe an asymmetric distribution (Adler et al., 2008; Caves, 2014). Impact factors vary significantly between disciplines and can depend on variables other than scientific quality, such as the number of authors per paper, article length, language, publication type, size of journal and the distribution of citations over time. There is no correction for self-citations—that is, articles citing other articles from the same journal—leading to 'coercive citation': the addition of extraneous citations to boost a journal's impact (Wilhite and Fong, 2012).

A journal's impact factor is, ostensibly, 'the average number of times an article published in the previous two years[8] was cited during the year in question' (Rossner et al., 2007). The reality is actually more complicated than that: rather than a simple mean, the numerator and denominator are measured from different popula-

[8] Clarivate also publishes a five-year impact factor

tions. Whereas the numerator includes all citations from journals indexed in the Web of Science, the denominator only considers *citable items*, which the company defines as follows.

> "Citable items are those items that comprise the figure in the denominator of the Journal Impact Factor calculation. These items are those identified in the Web of Science as an article, review or proceedings paper and are considered the substantive articles that contribute to the body of scholarship in a particular research field and those most likely to be cited by other articles. Other forms of journal content, such as editorial materials, letters, and meetings abstracts, are not considered as citable items."
>
> Clarivate Analytics (2019)

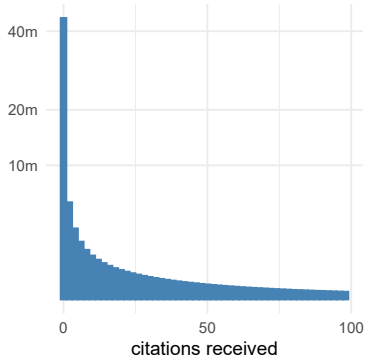


Figure 1.2: Histogram of citation counts to articles published in 2000–2010, according to data from Microsoft Academic

So to improve their impact factor—as well as merely trying to accrue more citations or publishing fewer articles—editors may also unscrupulously attempt to reclassify content as frontmatter, among other questionable tactics (PLoS Medicine Editors, 2006; Rossner et al., 2007; Falagas and Alexiou, 2008). Thomson Reuters argued around the same time however (McVeigh and Mann, 2009) that their classification process for citable items is 'accurate and consistent'. If the numerator of the calculation were restricted in the same way as the denominator, then impact factor scores would change considerably (Amin and Mabe, 2003).

Those issues aside, the distribution of citations received per article is skewed, as shown in Figure 1.2. This raises questions over the decision to summarise such a distribution with an arithmetic mean—or something a bit like one. Moreover, in the Journal Citation Reports, impact factors are published to three decimal places, which implies a spurious level of precision; most periodicals publish just a few dozen articles per year.

CiteScore was launched in December 2016 as a 'heavyweight rival' to Impact Factor (Van Noorden, 2016). At a high level, it uses essentially the same idea as Clarivate's metric: count the number of citations a journal has received and divide by the number of articles published. Key differences are that it collects citation data from the preceding three years (rather than two) and that the database used is *Scopus*, maintained by Elsevier, rather than Clarivate's *Web of Science*.

Some in the publishing industry have criticized Elsevier, an academic publisher, for producing a journal ranking metric, suggesting this is a conflict of interest (Van Noorden, 2016; Davis, 2016) however Elsevier disputed this argument and claim that Scopus metrics are more open and transparent, allowing researchers to audit whether any publisher has an unfair advantage or not (Straumsheim, 2016).

Bergstrom and West (2016)—whose Eigenfactor metric is published in Clarivate's *Journal Citation Reports*—did just that: comparing CiteScore with Impact Factor, they found that journals from Nature and Lancet Publishing Groups tend to receive lower CiteScores

relative to their Impact Factors. Their explanation for this phenomenon was that whereas Clarivate has a notion of 'citeable items' which excludes certain types of documents—such as news articles and editorials—from the denominator of the Impact Factor calculation, CiteScore counts all articles equally, research-based or not. Because publications such as *Nature* and *Science* contain a large number of such non-research items, their scores are 'highly diluted'.

On the other hand, one might make the reverse argument: that Impact Factor has been constructed in such a way that it gives an unfair advantage to publishers who choose to 'dilute' their journals in this way. Indeed, Bergstrom and West (2016) note: "by neglecting to count the front matter in its denominator, Impact Factor creates incentives for publishers to multiply their front matter". On the other hand, if the CiteScore overtook the Impact Factor in prominence, it would create a perverse incentive in the opposite direction, encouraging journals to reduce the amount of front matter, removing news articles, editorials and so on, to the detriment of their readership.

Some in the publishing industry appear to have had a rather visceral reaction to CiteScore, calling it 'quick, dirty and overtly biased' whilst claiming Impact Factor is 'a more reasonable metric' (Davis, 2016). No love is lost for *l'ancien régime* among academics, however. da Silva and Memon (2017) wrote: "... the JIF has always been an opaque marketing-based metric whose precise calculation was hidden behind a paywall, not making it verifiable by the public, and thus raising the ire of academics across the globe."

Google and Microsoft, as well as providing searchable databases with (greater or lesser) potential for collecting citation data (see Chapter 5), themselves publish some journal-level ranking metrics.

Google Scholar includes a 'Top publications'[9] page that ranks academic journals, both overall (by language) and in various broad categories and subcategories. It provides the disclaimer, "Dates and citation counts are estimated and are determined automatically by a computer program," though a measure of uncertainty is not given. Rankings are based on the h_5*-index*, which is a journal-level h-index for articles in a journal published in the last five calendar years (2014–2018 at the time of writing). The definition of h-index is the largest number h such that h articles have each received h or more citations. It was devised by Hirsch (2005) for evaluating authors, but has been adapted by Google for ranking publications. Shortly after its introduction, Google's h_5-index was mooted by some researchers (Baker, 2012) as a possible alternative to the journal impact factor.

Google Scholar's top 20 English-language journals by h_5-index in the subcategory 'Probability & Statistics with Applications', as of July 2019, are reproduced in Table 1.1.

[9] https://scholar.google.com/citations?view_op=top_venues&hl=en

An alternative Google-published metric is the 'h_5-median', which is the median number of citations received by the articles included in the h_5-index calculation.

Table 1.1: Top journals in the field 'Probability & Statistics with Applications' according to Google Scholar

Publication	h_5
Journal of Statistical Software	60
The Annals of Statistics	60
Journal of Econometrics	56
Journal of the American Statistical Association	53
Statistics in Medicine	49
Journal of the Royal Statistical Society: Series B (Statistical Methodology)	47
The Annals of Probability	45
The Annals of Applied Probability	40
Computational Statistics & Data Analysis	40
Probability Theory and Related Fields	39
Statistics and Computing	38
Journal of Computational and Graphical Statistics	35
Biometrika	34
Journal of Business & Economic Statistics	34
The Annals of Applied Statistics	34
Bernoulli	34
Mathematical Finance	34
The Stata Journal	32
Statistical Science	32
Biometrics	31

As with other statistics journal rankings that include citations from outside the field (e.g. Figure 3.14), more applied or interdisciplinary journals such as *Journal of Statistical Software* and *Statistics in Medicine* come highly ranked, and *Stata Journal* makes it into the top 20. However, the esteemed top three 'pure' statistical journals—the *Annals of Statistics*, *JRSS-B* and *JASA*—do well, too.

The journal *Econometrics*, possibly misclassified (depending on your personal view on whether econometrics fall within statistics) is highly ranked in statistics, but does not make it into the top twenty of the 'Business, Economics & Management' subcategory, if only because journals in that field tend to accrue citations more rapidly. Indeed, the 20$^{\text{th}}$-place business/economics publication, *Journal of Political Economy*, has an h_5-index of 73, greater than that of any probability or statistics journal.

Microsoft Academic[10] meanwhile ranks authors, institutions, journals and conferences by something it calls 'saliency', which is a weighted citation count, taking into account factors such as 'the reputation and the age of each citation'. Saliency is complemented by 'prestige', another metric, which equates to saliency per publication. This can be used when analysts wish to distinguish quality from quantity and avoid over-rating authors for being prolific, rather than consistently impactful. Other metrics such as citation count, h-index and publication count are also provided.

[10] https://academic.microsoft.com

Categories or 'topics' are algorithmically assigned by the Mi-

crosoft Academic Graph. The top 20 journals in 'statistics' (a topic within 'mathematics') according to Microsoft Academic are given below. The actual numerical values of saliency and prestige are not displayed in the Web interface, nor is any measure of uncertainty given.

Saliency

1. Journal of the American Statistical Association
2. Statistics in Medicine
3. Annals of Statistics
4. Biometrics
5. Journal of Econometrics
6. CA: A Cancer Journal for Clinicians
7. Biometrika
8. Computational Statistics & Data Analysis
9. Econometrica
10. Technometrics
11. Journal of Clinical Epidemiology
12. Journal of Statistical Software
13. Journal of The Royal Statistical Society Series B-statistical Methodology
14. Journal of Statistical Planning and Inference
15. IEEE Transactions on Signal Processing
16. Communications in Statistics-theory and Methods
17. Statistics & Probability Letters
18. Journal of Multivariate Analysis
19. PLOS ONE
20. Chemometrics and Intelligent Laboratory Systems

Prestige

1. CA: A Cancer Journal for Clinicians
2. Circulation
3. BMJ
4. Journal of Statistical Software
5. Journal of Machine Learning Research
6. Journal of The Royal Statistical Society Series B-statistical Methodology
7. Psychological Methods
8. Bioinformatics
9. Econometrica
10. Proceedings of the National Academy of Sciences of the United States of America
11. Systematic Biology
12. American Journal of Epidemiology
13. Journal of Clinical Epidemiology
14. NeuroImage
15. IEEE Transactions on Wireless Communications
16. Biostatistics
17. Oxford Bulletin of Economics and Statistics
18. Stata Journal
19. IEEE Communications Letters
20. Structural Equation Modeling

The prestige ranking especially leaves something to be desired, as it appears the statistics league table gets infiltrated by medicine,

biology and general science journals. This poses the question of how to define fields, which is discussed in greater detail in Chapter 4.

Despite the proliferation of competition, fewer citation metrics—not more—may be a better solution to the problem of trying to determine quality of research publications. Jeffrey Beall's famous list of predatory journals (discussed in the previous section) was accompanied by a list of 'misleading and fake metrics', which has, like the journal list, been continued by volunteers[11]. Even established metrics face strong criticism for their effect on the community; citation-based metrics from large bibliometric databases tend to favour traditional journals, possibly further entrenching the position of commercial publishers (Larivière et al., 2015). MacRoberts and MacRoberts (2017) make a similar criticism; as well as being 'unreliable', citation analysis may reinforce an 'elitist' view of science, tend to 'erase the contributions of women and minorities' and support the status quo.

[11] https://predatoryjournals.com/metrics

Even the creator of the impact factor, Garfield (1979) said, '... as with any methodology, citation analysis produces results whose validity is highly sensitive to the skill with which it is applied,' however he contended that metrics are 'a valid form of peer judgment that introduces a useful element of objectivity'. MacRoberts and MacRoberts (1996) disagree:

> "Today, in spite of an overwhelming body of evidence to the contrary, citation analysts continue to accept the traditional view of science as a privileged enterprise free of cultural bias and self-interest and accordingly continue to treat citations as if they were culture free measures... Neither of these assumptions is supported by the evidence."

1.5 Altmetrics

In an online era, a profileration of 'altmetrics' has emerged, based on such measures as the number of tweets or other social media mentions about an article, appearances in the news and on blogs, or how frequently an article has been viewed or downloaded from the publisher's web site or 'bookmarked' on services such as Mendeley. These alternative metrics—mostly at the article level—usually appear as numbers or badges beside works online, as a (sometimes weighted) count of how frequently the work has been mentioned in various media.

According to the site Altmetric[12]—one of several providers of alternative metrics; others being the Public Library of Science (PLoS) and the Elsevier's Social Science Research Network (SSRN)—these numbers are meant to complement, rather than replace, the older citation-based metrics. Some researchers (Fortunato et al., 2018) argue that altmetrics can solve some of the problems associated with citation-based metrics.

[12] https://www.altmetric.com

Advantages include being able to see the accumulation of 'attention' more quickly, rather than waiting several years for academic articles to be published, and being able to measure attention for a wider range of outputs that are not always conventionally cited, including data sets, software packages and presentations, as well as more traditional journal articles, conference proceedings and books.

Altmetrics are also not without their detractors. Colquhoun and Plested (2014) called web-based metrics 'childish' with 'ambiguous aims', adding: 'All bibliometrics give cause for concern, beyond their lack of utility. They do active harm to science'.

To consider another alternative: popular web sites such as Reddit[13] and StackOverflow[14] have voting systems that allow users to upvote or downvote posts and commentary based on whether they 'contribute to the discussion'. Low-quality, redundant or irrelevant content and comments are downvoted and become less visible, whilst better posts rise to the top. Whilst this may appear to work for sharing and discussing news or programming tips online, online commentary and feedback platforms for scientific work—on *BioMed Central*, *PLoS*, *BMJ* and *arXiv*—have 'failed to gain traction', possibly due to the smaller size of the academic community and the problem of anonymity (Neylon and Wu, 2009). There is however some support for moving away from a traditional journal-based system towards open-source formats such as *arXiv* and *PLoS One*, as these 'facilitate dissemination of new ideas and provide online realtime peer review for them' (Heckman and Moktan, 2018).

[13] https://www.reddit.com

[14] https://www.stackoverflow.com

1.6 Research assessment

In recent years, bibliometric indicators have been mooted as a possible way to 'add standardization to hiring, re-appointment, tenure and promotion decisions', albeit as a complement to rather than a replacement for traditional procedures (Holden et al., 2005). However, some authors believe citation metrics are being abused, with potentially serious repercussions. Indeed, Holden et al. (2006) followed up the aforementioned paper by explicitly recommending against using the impact factor as a proxy measure for researcher ability.

Caves (2014) coined *high-impact-factor syndrome* as the phenomenon among research institutions of using 'number of publications in high-impact-factor journals' to measure an academic's aptitude or potential. Heckman and Moktan (2018) found that having articles published in the 'top five' economics journals exerts an undue influence on whether or not a researcher receives tenure in an academic economics department in the US.

Excellence for Research in Australia (ERA)[15] is a national assessment that was announced by the Australian Government in 2008 and first took place in 2010. It is the counterpart to the UK Research Excellence Framework (REF)[16]. As part of this exercise, the Australian Research Council (ARC) produced a ranking of academic

[15] https://www.arc.gov.au/era

[16] https://www.ref.ac.uk

journals in each discipline, rating them in one of four categories: A*, A, B or C. Vanclay (2011) evaluated this ERA journal classification scheme and found it 'lacks sufficient rigour' and 'likely detrimental to several scientific disciplines', recommending to switch to an article-based approach instead. Following intense criticism, the A*–C rating scheme was discontinued for the 2012 round of the ERA (Mazzarol and Soutar, 2011), though the ARC continues to maintain a list of journals 'eligible for inclusion' as research outputs.

More recently, researchers in some Australian institutions have started looking to 'Q_1 journals'—those whose SCImago Journal Rank (SJR) is in the top 25% of the field.

In Chapter 6 we model the possibility that the UK's 2014 REF—ostensibly intended to be based on expert peer review only—was also implicitly employing a journal-based approach to rating institutions.

Whether or not impact factor abuse is part of the regular national assessments, it may nonetheless be institutionalized. Verma (2015) and Berenbaum (2019) describe how impact factors play a big role in performance reviews for academics: apparently some institutions will not even consider hiring someone who is not first author of a paper in a high-impact-factor journal, nor consider promoting someone until the average impact factor of the journals in which they have published meets a certain threshold.

Judging candidates by this metric—the venue of articles' publication versus those papers' actual citation counts—is a fundamental error that corrupts the results with significant aggregation bias. Even as an estimator for an author's citation counts, the mean is likely to over-estimate the citation impact of most papers (see Figure 1.2). As Caves (2014) points out: 'Giving extra credit for publications in HIF journals, i.e., for the company a paper kept, makes no sense… Just because a number is objective doesn't mean it is meaningful or informative.'

1.7 *Discussion*

Bibliometrics is a topic fraught with controversy. Actually reading a scientific paper will always be a better measure of its quality than trying to infer this from publication metadata such as citations. Nevertheless, not everyone has the time or technical expertise to do so.

Meanwhile, citation metrics are increasingly used—and abused—in the academic community, for want of a scalable and/or objective measure of research impact. This has a pernicious effect on scientific practice and places a lot of power in the hands of journal editors and publishers. It may not be helpful, however, to suggest abandoning citation metrics unilaterally, without providing pragmatic alternatives, as in the *p*-value debate (Wasserstein and Lazar, 2016).

In the face of this scenario, it is a task for scholars to ensure that bibliometrics is as 'statistical' as it claims to be, with greater reproducibility and better quantification of inherent uncertainty in

the calculation of various quantitative indicators.

In the following chapters we consider 'least worst' methods for measuring journal impact, be it the relative propensity for receiving citations within a field, the flow of influence between fields, how such fields are formed, or the role that journal identities have in formal research assessment—and the inherent uncertainty in estimating each of these effects. Such techniques should aim to minimize harm while remaining transparent and understandable to those who hope to use them.

2
PageRank and the Bradley–Terry model

This chapter discusses two alternative quantitative approaches for ranking scholarly journals, the Stigler (Bradley–Terry) model and the Eigenfactor (PageRank) score. Both of these methods offer advantages over the journal impact factor. The underlying models are described and their performance, both in principle and in practice, is compared. Each technique has been applied to cross-citation data from a sample of 47 statistical journals, as analysed in a recent paper by Varin et al. (2016). The benefits and pitfalls of such an analysis are discussed.

By examining the theory underpinning each method—one a generalized linear model and the other based on a Markov chain—it can be shown that the two are closely connected. Under idealized conditions, a modified PageRank metric—originally proposed by Pinski and Narin (1976), but mostly overlooked—yields rankings exactly equal to those from the Bradley–Terry model. We present a novel proof for this, and use the delta method to demonstrate some special cases where a PageRank-based score is an asymptotically efficient estimator for the Bradley–Terry model.

2.1 PageRank

PageRank, named for Larry Page, was developed in the 1990s by Google for their search engine (Page et al., 1999). It is a generalisation of the Pinski and Narin (1976) 'total influence' measure from bibliometrics. The connection between total influence and Markov chains was made by Geller (1978). For an interesting history, see Franceschet (2011) or Vigna (2016).

The Markov chain corresponding to PageRank has a simple analogy, of a random walk around the graph. Consider an imaginary PhD student, who opens a random journal from the library. The student selects at random a reference from within that journal and proceeds to read the cited journal. Then a third journal is selected from the references of the second, a fourth from the third, and so on. The proportion of overall time spent reading a particular journal may be seen as a measure of that journal's importance (Bergstrom, 2007).

Google's innovation[1] was the addition of a *damping factor*, α. At

[1] Strictly speaking, the damping factor was introduced as part of Katz centrality (1953). See Vigna (2016).

any time, with probability $1 - \alpha$, our random PhD student gets bored, returns their current journal to the shelf and selects a new journal at random from the library. In a random walk this would be the ability to 'teleport' randomly from one node to another. This helps link up disconnected components of sparse graphs. A PageRank computed with no chance of boredom/teleportation (i.e. with $\alpha \equiv 1$) is called *undamped* and is the same as the total influence metric of Pinski and Narin (1976).

To implement PageRank mathematically, the algorithm works as follows. Let $C = (c_{ij})_{n\times n}$ be a contingency table of citations (or Web hyperlinks, Twitter 'follows' or any other kind of directed connections), where c_{ij} is the number of citations to journal i from journal j. Set the diagonal of C to zero, to ignore journal self-citations. Compute the normalised (column-stochastic) matrix $\widetilde{C} = (\tilde{c}_{ij})_{n\times n}$, where $\tilde{c}_{ij} = c_{ij}/(\sum_{i=1}^{n} c_{ij})$.

Then PageRank is the stationary distribution of the Markov chain with transition matrix

$$P = \alpha\widetilde{C} + \frac{1-\alpha}{n}ee^T, \tag{2.1}$$

where e is an n-vector of ones. We can compute PageRank from the leading right eigenvector, $\pi = P\pi$. Undamped PageRank (total influence) is simply the stationary distribution of $\widetilde{C}$.

The *Eigenfactor* score was proposed by Carl Bergstrom (2007; 2008) and Jevin West (2010) as a (re)-adaptation of PageRank to bibliometrics. The Eigenfactor algorithm is an application of 'personalised' PageRank, where the transition matrix is defined

$$P_v = \alpha\widetilde{C} + \frac{1-\alpha}{n}ve^T, \tag{2.2}$$

with a *personalisation vector*, v, used to bias teleportation towards certain nodes (Franceschet, 2011). In the case of Eigenfactor, v is a vector of the number of articles published in each journal. In effect, this means the random PhD student selects an *article* at random at each step, then reads the containing journal. This introduces an explicit bias in favour of larger journals, which contain more articles.

The Eigenfactor score vector is derived from

$$\mathrm{EF} = \widetilde{C}\pi_v, \tag{2.3}$$

the product of the normalized citation matrix $\widetilde{C}$ and the personalised PageRank vector π_v, the latter being the stationary distribution of (2.2), $P_v\pi_v = \pi_v$ (West, 2010). This is an additional non-damped step (Vigna, 2016).

To control for the journal size bias (and distinguish 'prestige' from 'popularity') the *Article Influence* metric was introduced, defined as Eigenfactor score per article. That is,

$$\mathrm{AI}_i = \mathrm{EF}_i/v_i \tag{2.4}$$

for all $i = 1, \ldots, n$. An undamped equivalent was proposed by Pinski and Narin (1976) as the 'influence per publication'.

The Eigenfactor metrics published by Thomson Reuters (now Clarivate Analytics) use a constant damping factor of $\alpha = 0.85$, which is inherited from the value determined empirically by Google, but perhaps lacks theoretical backing to justify its re-application to citation networks (Newman, 2010).

2.2 *The Bradley–Terry model*

Stigler (1994) proposed a model that measured 'export scores' for academic journals. The analogy corresponds to bilateral trade; intellectual influence being 'exported' (i.e. citations received) and 'imported' (citations given) among trading partners (journals).

The system can be expressed as a Bradley–Terry paired comparisons model (Bradley and Terry, 1952); for two journals, i and j, assume that either i exports influence to (is cited by) j, or vice versa. The log-odds is given by

$$\text{log-odds}\,(i \text{ exports to } j \mid i \text{ and } j \text{ trade}) = \mu_i - \mu_j, \tag{2.5}$$

where μ_i and μ_j are the *export scores* of journals i and j, respectively. Journals can be ranked on a linear scale according to their export scores, where larger values imply greater influence (Stigler, 1994).

Estimates for the export scores can be computed via maximum likelihood estimation, using standard statistical software (Firth and Turner, 2012).

One problem with modelling citation data under a Bradley–Terry model, however, is that it assumes the citation counts are independently distributed, which can lead to overdispersion. To overcome this, Varin et al. (2016) describe the use of quasi-likelihood estimation (Wedderburn, 1974) to fit a 'quasi-Stigler' model,

$$\mathbb{E}(C_{ij}) = t_{ij}\pi_{ij} \tag{2.6}$$

$$\text{Var}(C_{ij}) = \phi t_{ij}\pi_{ij}(1 - \pi_{ij}), \tag{2.7}$$

where C_{ij} is the number of times journal j cites journal i; t_{ij} is the total number of citations between journals i and j; $\pi_{ij} = \text{logit}^{-1}(\mu_i - \mu_j)$, the probability that j cites i, using (2.5); and ϕ is the parameter of dispersion. For further details, see Varin et al. (2016).

2.3 *Comparison in principle*

Both journal ranking methods considered here are based on well-defined stochastic models. The Eigenfactor represents the stationary distribution of a Markov process, which has a simple interpretation as described earlier. The Stigler model is an example of the Bradley–Terry model—as used for ranking sports teams, for example—applied to pairs of 'competing' journals.

The Bradley-Terry model assumes independent matches (i.e. binomial trials) but this is not necessarily a valid assumption for journal citations—is it fair to say that references, particularly those from

within the same article's bibliography, are independent? Stigler (1994, § 9) acknowledges this potential criticism, but concludes the effects of within- and between-article dependencies seems to be mitigated by the aggregation of citations to the journal level. Furthermore, lack of fit can be investigated by analysis of the *journal residuals* (see § 2.4.3 below).

Invalid model assumptions are perhaps less of a problem for the Eigenfactor because it is not strictly a 'model' in the statistical sense; rather it is a unique characteristic of a matrix: an exact measure of the 'centrality' of the vertices in a graph. Nonetheless parameters may still be questioned: for instance, the Eigenfactor algorithm inherits PageRank's damping factor of $\alpha = 0.85$, determined empirically by Google when building their search engine, but possibly without much theoretical backing (Newman, 2010, § 7.4). Is it appropriate to assume a 'random surfer' visiting pages on the Web will behave in the same way as our imaginary PhD student randomly reading academic journals? Perhaps a different value of α might be more effective.

Another possible criticism of the Eigenfactor is that, although self-citations might be subject to manipulation by unscrupulous journal editors (Wilhite and Fong, 2012), they are not always irrelevant or misleading: some journals may specialize in particular fields or sub-fields, in which case authors would genuinely refer to many articles from the same journal. Much of the citation data from such articles will be ignored by the Stigler model and the Eigenfactor algorithm, which both disregard self-citations entirely. The impact factor, meanwhile, counts self-citations with exactly the same weight as citations to other journals. Metrics need not treat self-citations on an all-or-nothing basis: the SJR indicator (González-Pereira et al., 2010) considers self-citations, but limits their number to 33% of an article's total references.

Publication size bias—i.e. bigger journals getting higher scores irrespective of quality—is another concern for journal ranking metrics. The Stigler model does not explicitly take the size of each journal (i.e. the total number of articles published) into account. Stigler (1994, § 8) showed, however, that bias based on 'size' alone is absent, by using the following simple argument.

Consider three journals, A', A'' and B. What if we merge A' and A'' into a single journal, $A = A' \cup A''$? Let O' represent the odds that journal A' cites B, and let O'' be the odds that journal A'' cites B. Then the corresponding odds for A are an appropriately

weighted average of O' and O'':

$$\begin{aligned} O &= \frac{P(A \text{ cites } B)}{P(B \text{ cites } A)} \\ &= \frac{P(A' \text{ cites } B) + P(A'' \text{ cites } B)}{P(B \text{ cites } A)} \\ &= \frac{P(B \text{ cites } A')\dfrac{P(A' \text{ cites } B)}{P(B \text{ cites } A')} + P(B \text{ cites } A'')\dfrac{P(A'' \text{ cites } B)}{P(B \text{ cites } A'')}}{P(B \text{ cites } A)} \\ &= \lambda O' + (1-\lambda)O''. \end{aligned}$$

A similar result for *invariance to splitting of journals* exists for eigenvector centrality measures (Palacios-Huerta and Volij, 2004) and might reasonably be applied to the Eigenfactor (Franceschet, 2010).

On the other hand, it is important to note that such proofs relating to splitting or merging of journals are concerned with the merging and splitting of link weights (citation counts), overlooking the measure more likely to be used for the 'size' of a journal: the total numbers of articles published, which are separate data. Indeed, the creators of the Eigenfactor acknowledge that the raw Eigenfactor score tends to be larger as article count increases, and that this represents the greater importance of the larger journals (Bergstrom, 2007). To evaluate an article's influence based on 'the company it keeps' rather than the size of the journal in which it was published, a separate Article Influence score may be computed from the Eigenfactor via an explicit normalization step (2.4).

The Journal Impact Factor is a global metric that depends on citation counts but not necessarily on the field of the journal or of those that cite it—though refinements have been proposed, such as field-normalized impact factors (Leydesdorff et al., 2013). The Stigler model only considers citations between the journals in the list being modelled (Varin et al., 2016, p.16), thus generating a localised ranking within a field that ignores journals' potential influence in the wider academic community. In principle the Eigenfactor is a 'global' ranking for the whole Web of Science, but the algorithm may be applied to only a subset of journals instead, as demonstrated in the following section.

More generally, the arbitrariness of using an algorithm like the Eigenfactor may itself be questioned. Arguably, the random surfer or PhD student are analogies to aid explanation rather than real world phenomena—or representations thereof—that researchers are trying to model. In that sense, the rules of an algorithm are themselves arbitrary and there is nothing special about the output of one algorithm over another. Unlike a statistical model fit to observed data, for Eigenfactor there is no notion of lack of fit, nor the possibility of expanding the model or measuring its precision. In that sense, the algorithm as a whole does not bear scrutiny, since there is no way of saying how well it represents the world, if indeed it represents anything at all.

2.4 Comparison in practice

The quantitative analyses in this section are based on cross-citation data for 47 statistics and probability journals, retrieved from Thomson Reuters' 2010 edition Journal Citation Reports by Varin et al. (2016). The results for the Stigler model were fitted using code based on the supplementary files provided with that paper.

The Eigenfactor algorithm, as described in the appendices of West (2010), was implemented in R (R Core Team, 2019) and is available in the `scrooge` package on GitHub[2]. The article counts vector, a, was not explicitly provided in the data accompanying Varin et al. (2016). For this chapter, the article counts for each of the 47 journals were manually retrieved from the 2010 edition of the Web of Science.

[2] `https://github.com/Selbosh/scrooge`

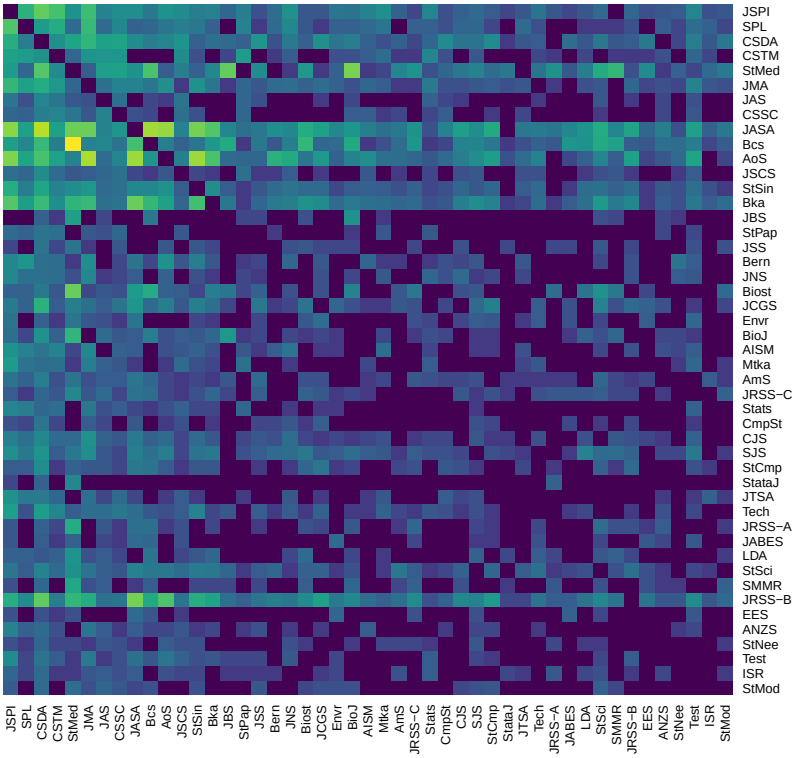
Figure 2.1 gives visualizations of the journal network, by raw citation counts and scaled by the total number of references in the citing journal, both with self-citations omitted. The matrix is quite sparse. Notice that many of the 'largest' journals receive a high number of citations from across the network. It is also apparent that *The Journal of the Royal Statistical Society, Series B* (JRSS-B) receives a relatively high number of citations for its size.

2.4.1 Scores and rankings

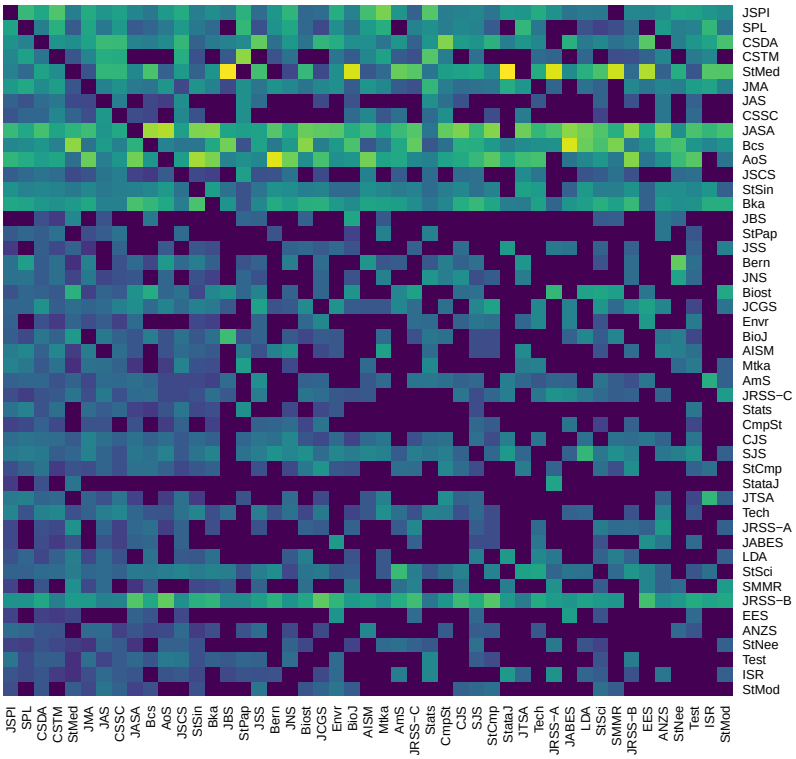
Stigler-model export scores were estimated using the `scrooge` package and verified using the `BradleyTerry2` package (Firth and Turner, 2012). A centipede plot of the Stigler-model export scores is given in Figure 2.2. The ranking matches that originally produced by Varin et al. (2016) using the same data. We find that the *Journal of the Royal Statistical Society, Series B* is a clear leader, followed by a small group comprising the *Annals of Statistics* (AoS), *Biometrika* (Bka) and the *Journal of the American Statistical Association* (JASA). The *Journal of Applied Statistics* (JAS) is bottom of the ranking. Interpretation of the 'comparison intervals' is discussed in the next subsection.

Eigenfactors and Article Influence scores were computed in R and are plotted in Figures 2.3a and 2.3b, respectively. Both scores give the top four places to the same top four journals identified by the Stigler model, though in slightly different orders. These four journals—JRSS-B, AoS, *Biometrika* and JASA—are generally considered the most prestigious in statistics (Varin et al., 2016) and so it is reassuring that all three metrics place them top of the field. Which journals occupy the *bottom* of the league is not the same across the three rankings, but this is perhaps less important.

Figure 2.4a shows that Eigenfactor scores tend to increase with journal size. The two (log-transformed) variables have a moderately positive correlation, with Pearson correlation coefficient 0.5. Indeed, JRSS-B is not the leading journal if ranked by Eigenfactor, but notice that every journal with a higher Eigenfactor score than JRSS-B also



(a) Raw counts, C



(b) Scaled by column, $\widetilde{C}$

Figure 2.1: Heat maps of the 47×47 journal cross-citation matrix from 2010 JCR data. Journals are sorted in descending order by article count. Lighter squares represent more citations from the column journal to the row journal

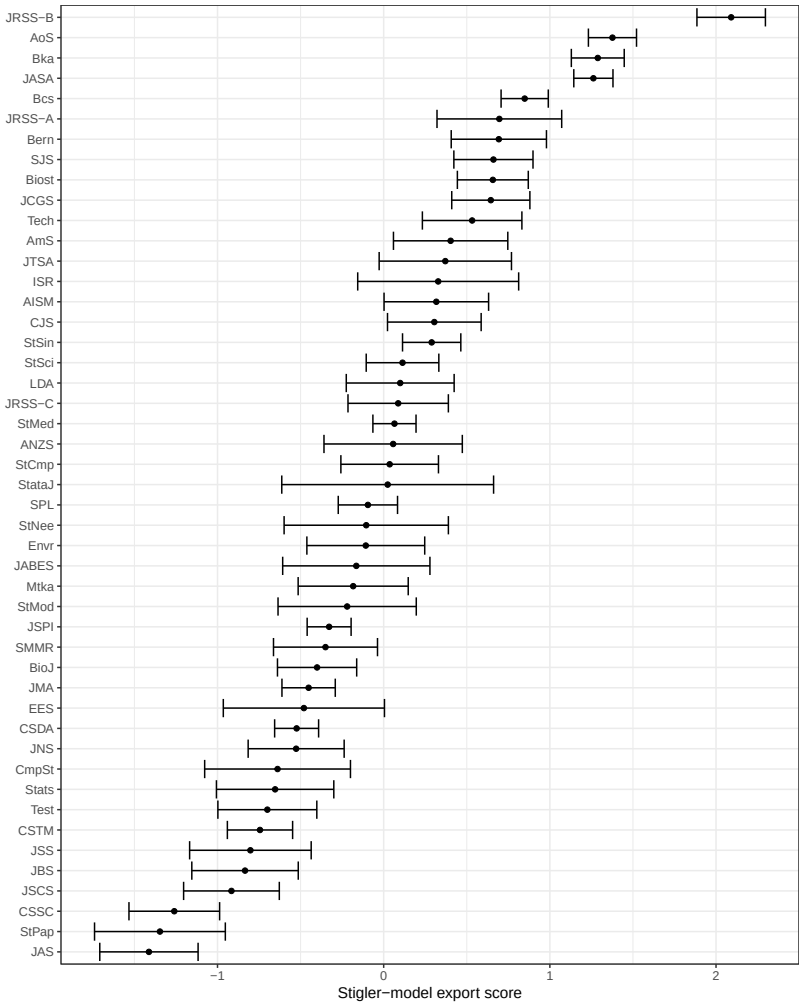
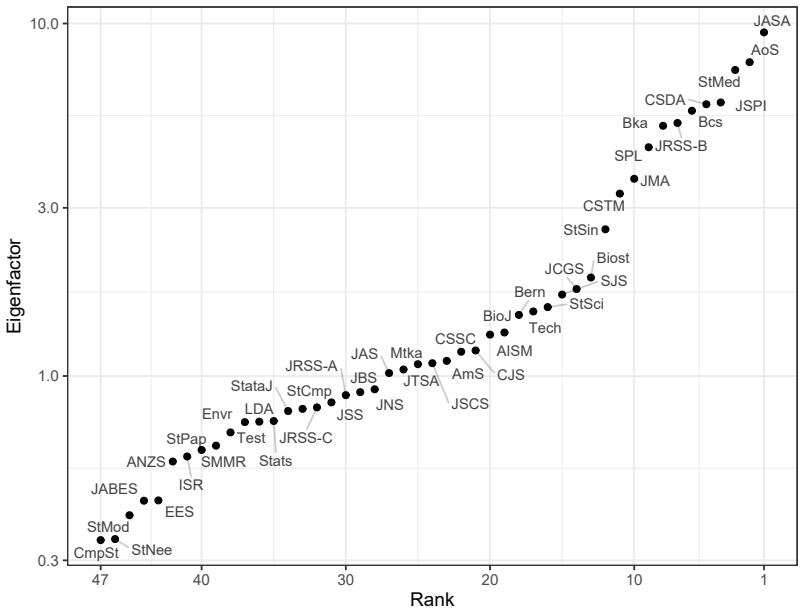
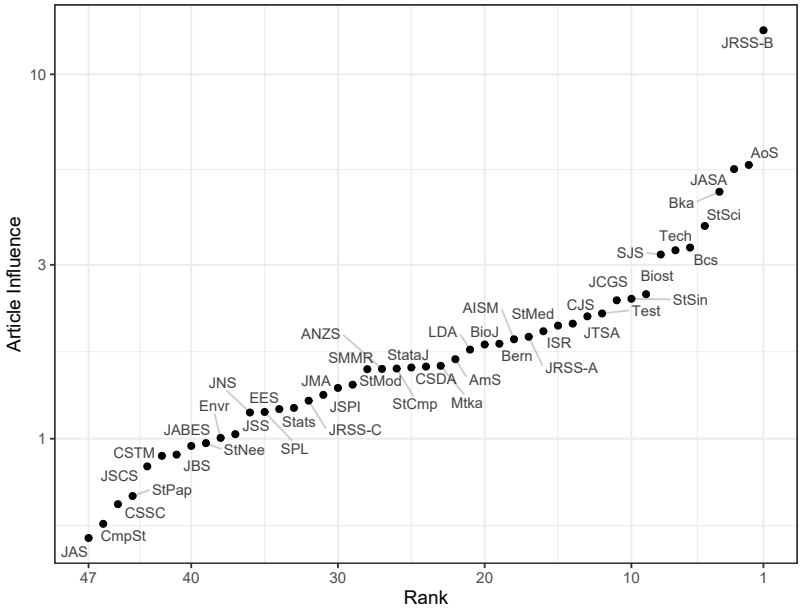


Figure 2.2: Centipede plot of estimated journal export scores and 95% 'comparison intervals' (Firth and de Menezes, 2004) for 2010 JCR data. The points represent estimated journal export scores; their error bars correspond to $\pm 1.96 \times$ quasi-standard-error of each score

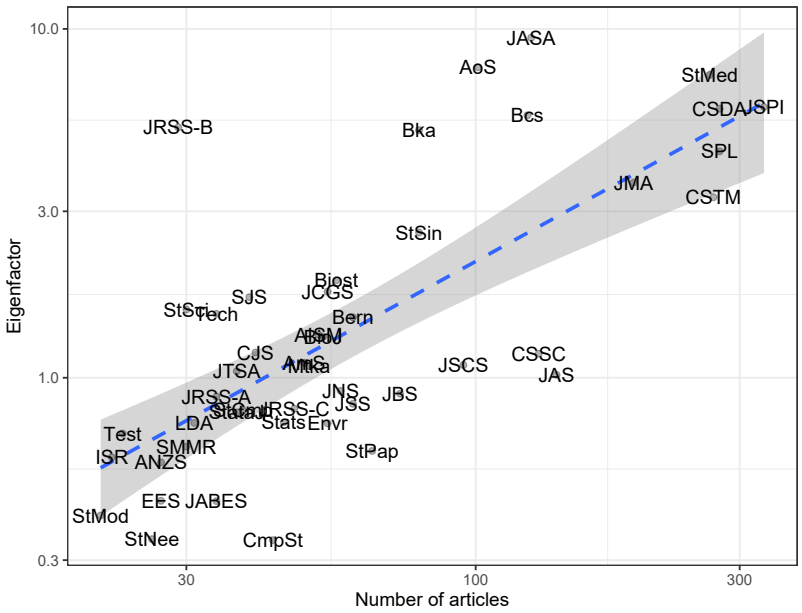


(a) Eigenfactor scores

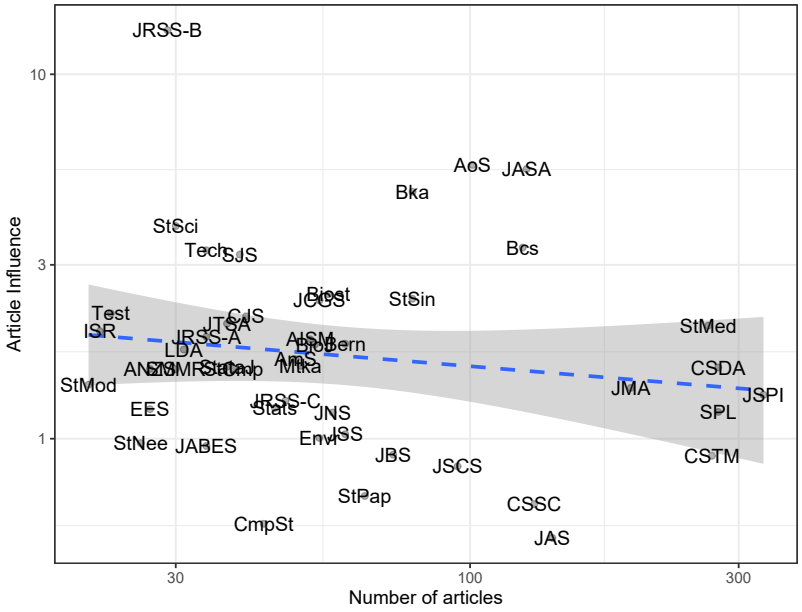


(b) Article Influence scores

Figure 2.3: Distribution of sorted Eigenfactor and Article Influence scores for statistics journals from 2010 JCR data. Journals are labelled and scores are on a logarithmic scale



(a) Eigenfactor



(b) Article Influence

Figure 2.4: Scatter plots showing Eigenfactor metrics against journal size, on a log-log scale, with lines of best fit and 95% confidence bands

has considerably more articles.

Compare Figure 2.4b, depicting the Article Influence score against article count. The (log-log) correlation between Article Influence and journal size is weaker, with Pearson correlation coefficient -0.25. In a ranking of Article Influence scores, JRSS-B retakes its crown, followed by *Biometrika*, *Annals of Statistics* and the *Journal of the American Statistical Association*.

A comparison of rankings is given in Figure 2.5. Many journals are ranked similarly between the Stigler model and Article Influence scores; a few move up or down several places, while a small minority experience dramatic changes in rank. For example, notice that *Test* was ranked 10$^{\text{th}}$ by Article Influence, but 40$^{\text{th}}$ by the Stigler model. Conversely, *Stata Journal* (StataJ) was a lowly 43$^{\text{rd}}$ for Article Influence but a middling 24$^{\text{th}}$ according to the Stigler model.

A plot of Article Influence scores against Stigler-model export scores is given in Figure 2.6. There is a strong positive relationship: the Pearson correlation coefficient between the log-transformed Article Influence score and the Stigler export score is 0.84. The Spearman rank correlation coefficient is 0.85. Journals that saw big shifts in ranking in Figure 2.5 are easily identified here: *Stata Journal*, *Test* and *Statistical Science* (StSci) appear to be outliers. One possible explanation is that these three publications are relatively small. In addition, *Stata Journal* is unusual, being a publication dedicated to a proprietary software program.

2.4.2 *Estimation uncertainty*

Some authors argue (Varin et al., 2016) that so-called 'statistical' bibliometric journal rankings should not be considered as such without some measure of estimation uncertainity. By this principle, the Stigler model is a statistical ranking method but journal impact factors, PageRanks and the Eigenfactor metrics are not.

Hypothesis tests for the Bradley–Terry model's estimated 'ability scores'—and hence Stigler-model 'export scores'—are well defined (Bradley and Terry, 1952). It is therefore possible to construct tests for the significance of differences in pairs of export scores, e.g. testing if $\mu_i - \mu_j = 0$ for journals i and j. From this we can deduce whether it is appropriate to say that journal i is really ranked higher than journal j (and vice versa) or if it is too close to call.

Conventionally, to compare any such pair of parameters would require the full variance-covariance matrix of the estimates $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, which would be cumbersome and impractical to print even for a relatively small number of journals. Firth and de Menezes (2004) propose the use of *quasi-variances* as an economical alternative, with

$$\operatorname{Var}(\hat{\mu}_i - \hat{\mu}_j) \approx \operatorname{qvar}_i + \operatorname{qvar}_j, \tag{2.8}$$

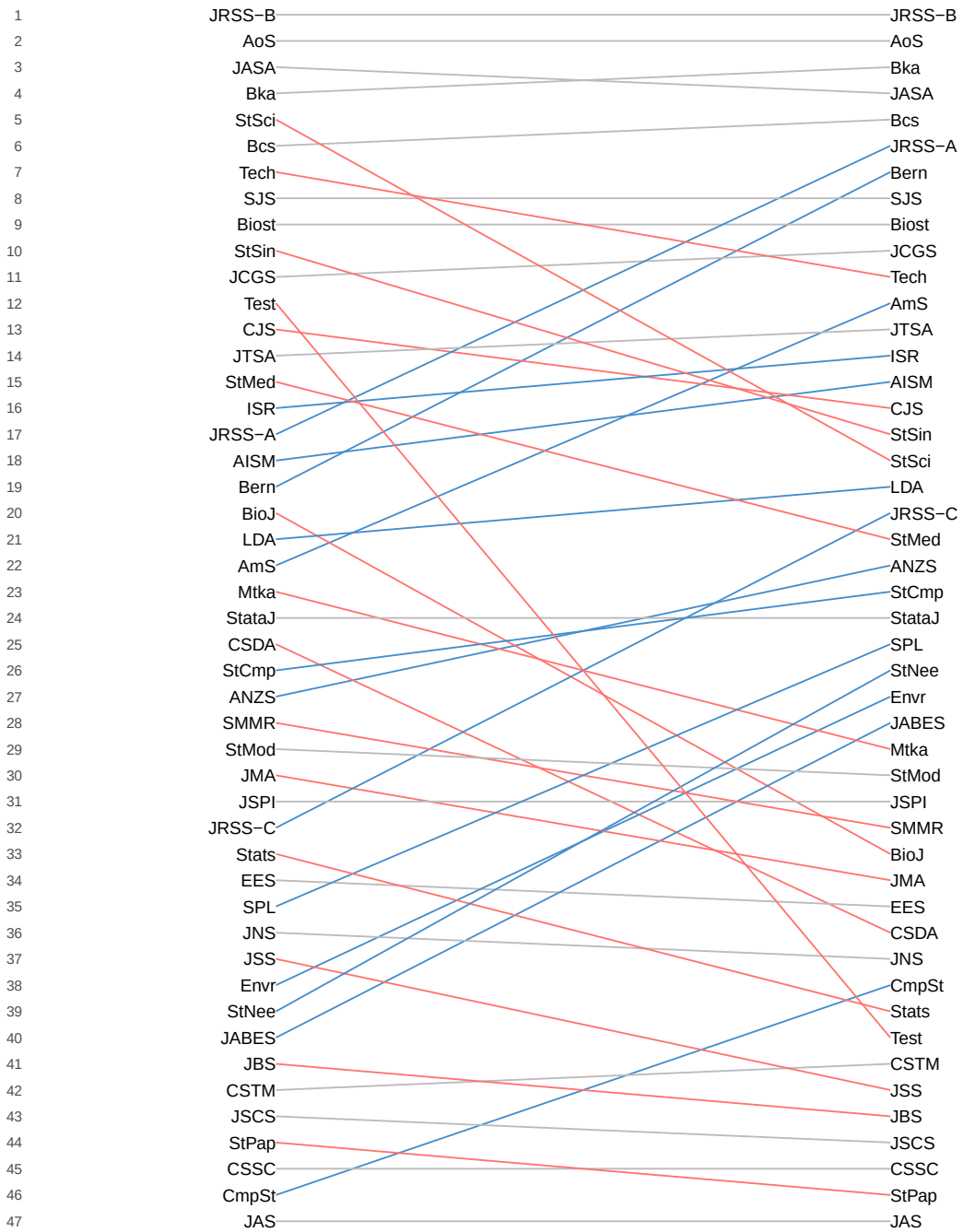


Figure 2.5: Comparison of journal rankings by Article Influence score and by Stigler-model export score. Red lines denote journals ranked more highly by Article Influence, blue lines denote journals ranked more highly by the Stigler model and grey lines denote journals ranked equally (± 1 place) by the two methods

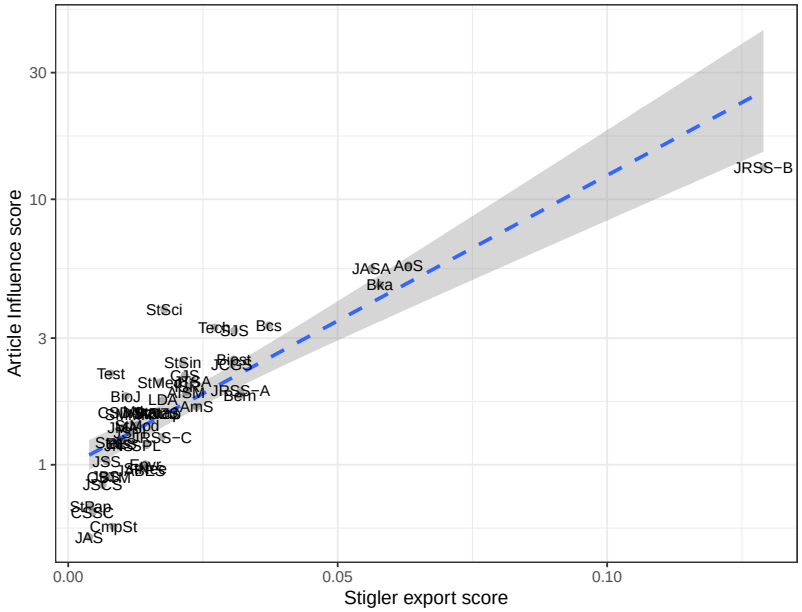


Figure 2.6: Scatter plot of Article Influence score (on a log-scale) against estimated Stigler export score, with a line of best fit

where the quasi-variances are found by minimising

$$\sum_{i<j} p\left(\text{qvar}_i + \text{qvar}_j, \text{Var}(\hat{\mu}_i - \hat{\mu}_j)\right) \tag{2.9}$$

for some penalty function $p(x, y) \geq 0$. In this case we used the squared log difference,

$$p(x, y) := (\log x - \log y)^2. \tag{2.10}$$

Inexactness of a quasi-variance approximation may be summarised by the relative errors of quasi-standard-errors from their corresponding standard errors derived from the full variance-covariance matrix. Firth and de Menezes (2004) suggested reporting the 'worst' relative errors and the `qvcalc` package includes these in the summary output. The distribution of relative errors, not just their minimum and maximum, may be a better indicator of problems with quality of a quasi-variance approximation, as shown by Figure 2.7.

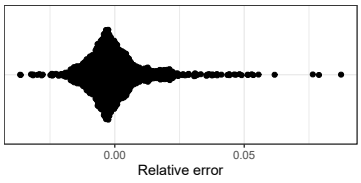


Figure 2.7: Bee swarm plot showing the distribution of relative errors of the quasi-variance approximation to simple contrasts in the fitted Stigler model

The Eigenfactor algorithm as described in West (2010) does not provide any measure of uncertainty for the computed Eigenfactor or Article Influence scores.

The data for a given year can be thought of as estimating an underlying vector of latent variables corresponding to journal 'states', which change over time. So although the citation counts data are (arguably) exact and complete, the observable citations are only a manifestation of the hidden variables that describe the true underlying journal 'quality'.

Considering the ease of constructing hypothesis tests for the Stigler model, it would be desirable therefore to try and derive some form of 'error bars' for the Eigenfactor.

Rosvall and Bergstrom (2010) suggested that resampling techniques such as the bootstrap (Efron, 1979) can be applied to anal-

yses of citation networks. Resampling the nodes (i.e. the journals), however, is not a good idea: for example, how does one make sense of a citation network with two *Biometrikas*?

Alternatively, one can consider resampling the link weights, i.e. the citation counts, using a non-parametric approach. Treat each citation as an independent event. A delete-m jackknife might involve all possible subsets of $N - m$ citations (Davison and Hinkley, 1997, § 2.8) but finding all $\binom{N}{m}$ of them proves computationally infeasible even for small m. It may be enough to select R such subsets at random. Construct the corresponding citation matrices $\mathbf{C}_{1:R}$ and each time compute Eigenfactor scores for each journal using the same procedure from § 2.1. The delete-m jackknife estimate of the variance of an Eigenfactor score is then

$$\mathrm{Var}_{R,m}(\mathrm{EF}) = \frac{N-m}{m}\frac{1}{R}\sum_{r=1}^{R}\left(\mathrm{EF}_r - \mathrm{EF}\right)^2, \tag{2.11}$$

where EF_r and EF are the Eigenfactor scores computed from the subsampled and complete citation data, respectively. The process would be equivalent for Article Influence scores.

Unfortunately, despite an apparently sound theoretical framework relating sampled variances to the 'true' Eigenfactor uncertainty, this method relies on assuming citations are all independent, which surely they are not, especially within the bibliography of a single published article.

Eldardiry and Neville (2008) argue that resampling techniques assuming independent, identically-distributed observations 'consistently underestimate the variance of sampling distributions in relational data' because dependencies among the observations reduce the *effective sample size* of the data. The sample size may be spuriously increased for any network with large numbers of citations, even if the number of journals remained the same.

One might instead consider a parametric bootstrap, where each link weight (citation count) is sampled parametrically from independent Poisson distributions with the observed link weights as their means (Rosvall and Bergstrom, 2010). This approach is still an oversimplification because it assumes independence of citation counts. In addition, any Poisson resampling is performed before the column-standardization step in the Eigenfactor algorithm—if our sampler generated a column of 1s and and a column of 100s, these apparently very different vectors would be equal after column-scaling. This may have the effect of reducing the variability of samples.

Mirshahvalad et al. (2013) showed that Poisson citation resampling underestimates the variance of the link weights, but that a more sophisticated parametric bootstrap using multinomial sampling requires article-level data. Where the latter are unavailable, they proposed a 'minimal model', using parametric resampling sampling from the distribution

$$\mathrm{Pois}(N_1) + 2\mathrm{Pois}(N_2), \tag{2.12}$$

where $N_1 = 2w^{0.9} - w$ and $N_2 = w - w^{0.9}$, for each observed citation count w, so that the variance of each link weight is

$$\begin{aligned} \mathrm{Var}\{\mathrm{Pois}(N_1) + 2\mathrm{Pois}(N_2)\} &= 2w^{0.9} - w + 4(w - w^{0.9}) \\ &= 3w - 2w^{0.9}. \end{aligned} \tag{2.13}$$

However it is difficult to assess the validity of model (2.12)—which was proposed for use in clustering, rather than computation of centrality measures—without access to article-level data or some other external validity criterion.

One solution designed specifically to deal with dependencies in network data is *relational subgraph resampling*, proposed by Eldardiry and Neville (2008). However, we will not implement it here.

For this chapter we consider a parametric multinomial resampler, drawing from a multinomial distribution with $N = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}$ independent trials, n^2 categories (one per entry in the citation matrix) and probabilities $p_{ij} = c_{ij}/N$ for $i, j = 1, \ldots, n$. Draw R samples from this distribution and compute their respective Eigenfactor and Article Influence scores. Then the estimated variance is simply the sample variance

$$\mathrm{Var}_R(\mathrm{EF}) = \frac{1}{R}\sum_{r=1}^{R}(\mathrm{EF}_r - \mathrm{EF})^2, \tag{2.14}$$

and confidence intervals for the scores can be expressed in the form

$$\mathrm{EF} \pm 1.96\sqrt{\mathrm{Var}_R(\mathrm{EF})} \tag{2.15}$$

for each journal.

Figures 2.8 and 2.9 provide centipede plots of Eigenfactor and Article Influence scores. A logarithmic transformation has been applied for ease of comparison of journal rankings. There is a noticeable size effect in the widths of the confidence intervals: the larger the Eigenfactor score, the wider the interval. (However this effect appears reversed in Figure 2.8 due to the log transformation.)

Confidence intervals in Figure 2.9 can be compared directly with the Stigler-model comparison intervals of Figure 2.2. The resampled intervals seem narrower, which might corroborate earlier arguments for resampling methods underestimating the variance. Nonetheless most consecutive journals have overlapping confidence intervals and do not have significantly different scores. This implies—as in the Stigler model—that small differences in Article Influence scores should not be over-interpreted.

Smaller journals have larger Article Influence standard errors, possibly due to their lower citation counts: *Stata Journal* could place anywhere in the bottom 10 ranks, while *Test* ranges from 6$^{\text{th}}$ to 17$^{\text{th}}$. At the top of the table, the inferences are similar: JRSS-B is significantly ahead of all of the other journals, while JASA, AoS and *Biometrika* are mutually indistinguishable. It also seems that *Statistical Science* (StSci) has a similar Article Influence score to *Biometrika*—despite the former having a fairly middling Stigler-model score.

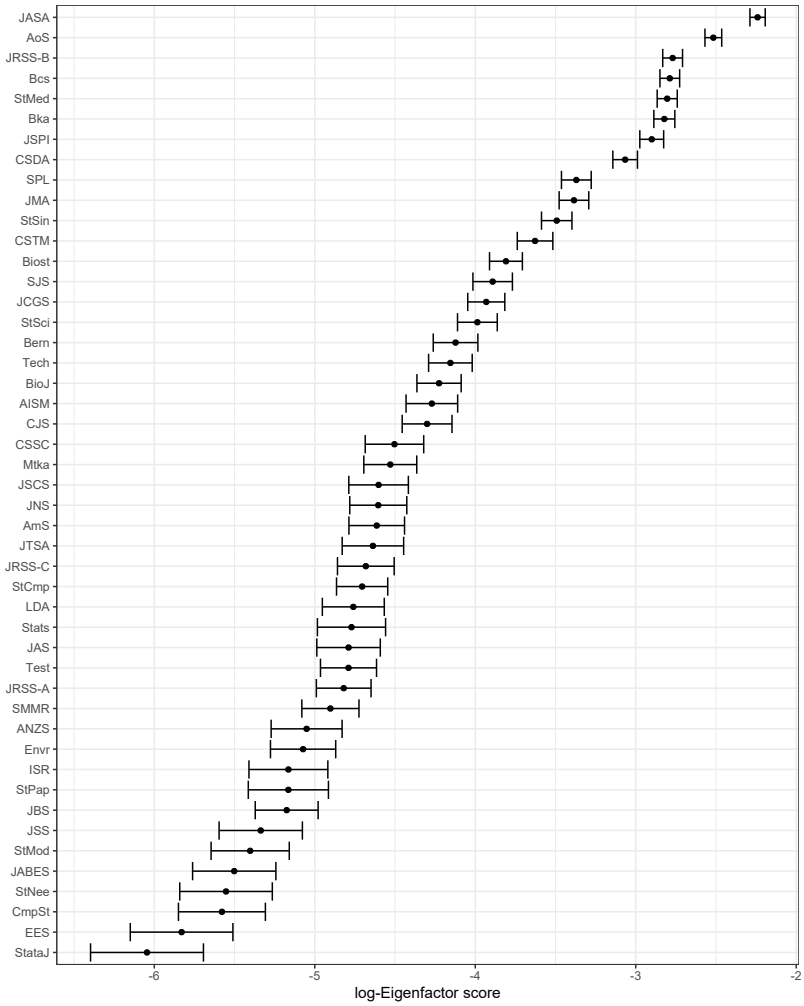


Figure 2.8: Centipede plot of log-Eigenfactor scores and 95% confidence intervals, based on multinomial resampling of 2010 JCR data with 500 replications

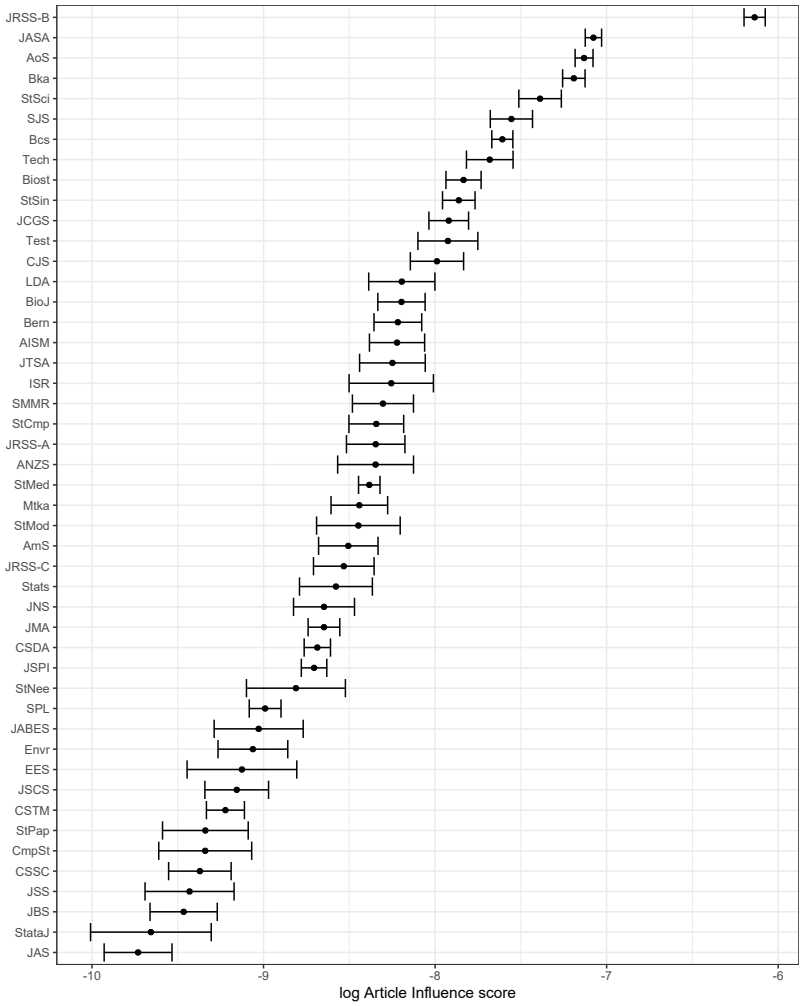


Figure 2.9: Centipede plot of log-Eigenfactor scores and 95% confidence intervals, based on multinomial resampling of 2010 JCR data with 500 replications

2.4.3 *Lack of fit; validity*

Though it is interesting to compare the Stigler and the Eigenfactor-based rankings to each other, such an exercise is not valuable if neither of the underlying methods fits or explains the data well.

Goodness of fit for the Stigler model may be assessed through the analysis of *journal residuals*, defined by Varin et al. (2016) to be

$$r_i = \frac{\sum_{j=1}^{n} \hat{\mu}_j r_{ij}}{\sqrt{\hat{\phi} \sum_{j=1}^{n} \hat{\mu}_j^2}}, \tag{2.16}$$

where r_i denotes journal residual for journal i, $\hat{\mu}_1, \ldots, \hat{\mu}_n$ are the estimated export scores, r_{ij} is the Pearson residual for citations of journal i by journal j and $\hat{\phi}$ is the estimated parameter of overdispersion. The journal residuals should, under the Stigler model, be normally distributed and uncorrelated with the export scores.

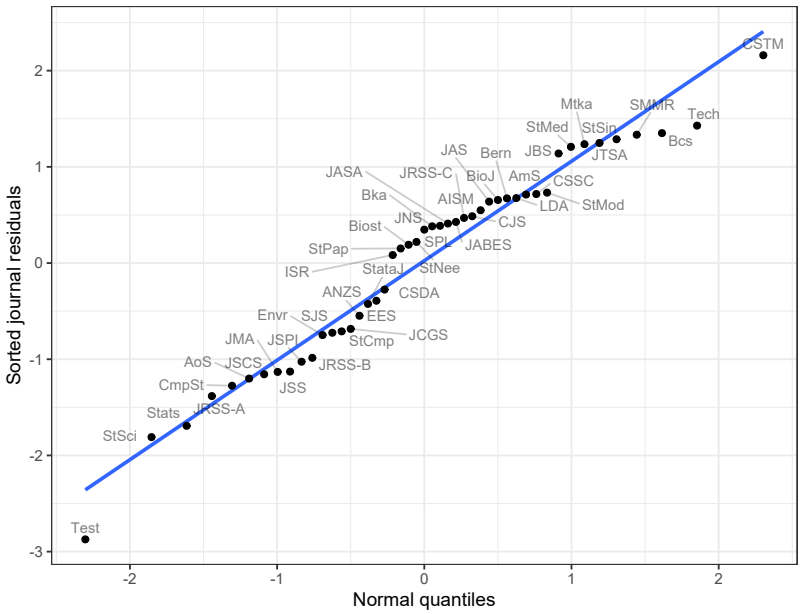


Figure 2.10: Normal Q-Q plot of journal residuals for the fitted Stigler model

A normal Q-Q plot of journal residuals is given in Figure 2.10. A plot of journal residuals against fitted values is provided in Figure 2.11. From these visualisations, it appears that the journal residuals are approximately normally distribution and uncorrelated with the fitted Stigler model export scores. There are no distinctive outliers. From this we can infer that the data do not violate these assumptions of the model.

Another way to measure the lack of fit of a log-multiplicative model is the index of dissimilarity (Kuha and Firth, 2011). The index, $\hat{\Delta}$ represents the estimated proportion of counts that must be moved from one cell to another in the expected contingency table to yield exactly the observed counts. It is given by the formula

$$\hat{\Delta} = \frac{\sum_{i \neq j} |c_{ij} - \hat{c}_{ij}|}{2 \sum_{i \neq j} c_{ij}}, \tag{2.17}$$

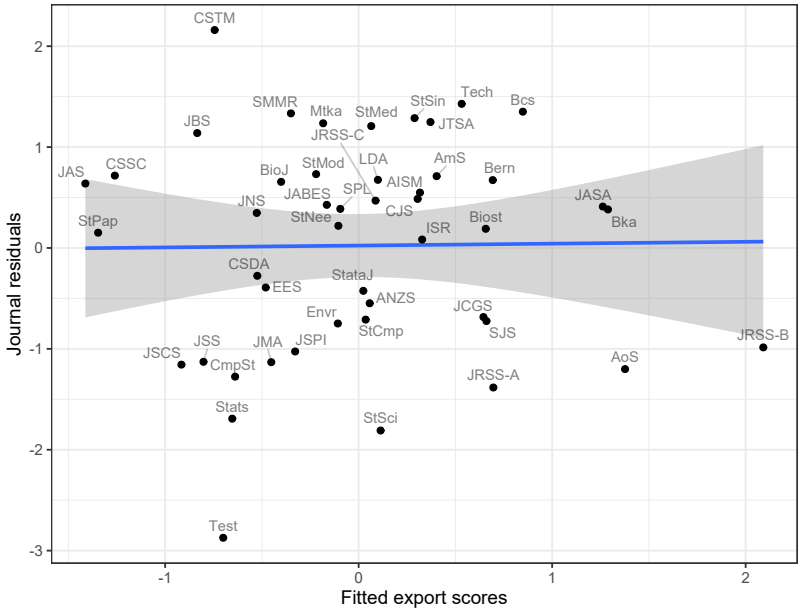


Figure 2.11: Journal residuals against export scores for the fitted Stigler model

where in this case $\hat{c}_{ij}$ is the expected number of citations from journal j to journal i under the fitted Stigler model. This number could be employed to compare two competing models, but may not be very useful on its own.

The validity of the Eigenfactor is harder to evaluate. Its underlying algorithm is not a 'model' in the statistical sense and makes no predictions that might be internally validated via residual analysis. At a high level, one might question some of the assumptions behind the Eigenfactor—as in § 2.3—but it is not straightforward to check quantitatively the validity of computed Eigenfactor and Article Influence scores, beyond computing their correlation with alternative ranking methods (e.g. Research Assessment Exercise results, as in § 6 of Varin et al., 2016, and in Chapter 6 of this thesis).

2.5 *Theoretical connection*

To motivate the theoretical relationship to be shown between the Bradley–Terry model and PageRank/Eigenfactor, we use a little bit of linear algebra and of Markov chain theory.

2.5.1 *Quasi-symmetry*

The Bradley–Terry model is a logistic formulation of the quasi-symmetry model (Agresti, 2013, Chapter 10). Quasi-symmetry was originally defined by Caussinus (1965): an $n \times n$ matrix C is called *quasi-symmetric* if its elements can be decomposed in the form

$$c_{ij} = \alpha_i \beta_j \gamma_{ij}, \tag{2.18}$$

where $\gamma_{ij} = \gamma_{ji}$. From (2.18) we can derive a simpler decomposition

$$c_{ij} = d_i s_{ij}, \tag{2.19}$$

where $d_i = \alpha_i / \beta_i$ and $s_{ij} = s_{ji} = \beta_i \beta_j \gamma_{ij}$. In matrix form, this relationship is represented by

$$C = DS, \tag{2.20}$$

where D is a diagonal matrix and S is symmetric. The Bradley–Terry model attempts to retrieve the ability scores $\mu_i \equiv \log d_i$ for all $i = 1, \ldots, n$. In the linear algebra literature, quasi-symmetric matrices are called *symmetrizable* and the matrix D^{-1} is called the *symmetrizer* (Dias et al., 2016).

From (2.18) and (2.19) it is easy to show that a quasi-symmetric matrix satisfies the property

$$c_{ij} c_{jk} c_{ki} = c_{ji} c_{kj} c_{ik} \tag{2.21}$$

for each triplet $i, j, k = 1, \ldots, n$ (Caussinus, 1965; Sharp and Markham, 2000).

What, then, is the connection with PageRank? Consider the transition matrix of a Markov chain. For brevity of notation we will look at the transition matrix of a discrete-time Markov chain; equivalent results hold for rate matrices of continuous-time Markov chains. Given a (column-stochastic) transition matrix P, there exists a stationary distribution, $\pi = (\pi_1, \ldots, \pi_n)$, if *global balance*,

$$\pi_i \sum_{j=1}^{n} p_{ji} = \sum_{j=1}^{n} \pi_j p_{ij}, \tag{2.22}$$

is satisfied for all $i = 1, \ldots, n$. The detailed balance equations (also known as local balance)

$$\pi_i p_{ji} = \pi_j p_{ij}, \tag{2.23}$$

hold for all pairs $i, j = 1, \ldots, n$ if and only if the Markov chain is reversible.

Kolmogorov's criterion, which is easy to derive from (2.23), is that

$$p_{ij} p_{jk} p_{ki} = p_{ji} p_{kj} p_{ik} \tag{2.24}$$

for every triplet (i, j, k) if and only if the Markov chain is reversible (Kelly, 1979, Chapter 1).

The similarity between (2.21) and (2.24) is striking. In fact, it follows that a Markov chain is reversible if and only if its probability transition matrix is quasi-symmetric (McCullagh, 1982; Bof et al., 2017).

2.5.2 *The Scroogefactor*

'Oh! But he was a tight-fisted hand at the grindstone, Scrooge! a squeezing, wrenching, grasping, scraping, clutching, covetous, old sinner! Hard and sharp as flint, from which no steel had ever struck out generous fire; secret, and self-contained, and solitary as an oyster.'

Dickens (1843)

Pinski and Narin (1976) proposed three different journal ranking metrics: influence weight, total influence and influence per publication. The *influence weight* for a journal, w_i, is defined by the recursive equation

$$w_i = \frac{\sum_{k=1}^{n} w_k c_{ik}}{\sum_{j=1}^{n} c_{ji}} \tag{2.25}$$

for each journal $i = 1, \ldots, n$, where c_{ij} is the number of citations journal i receives from journal j. Hence, *influence per publication* is defined as influence weight multiplied by the number of (outgoing) references per article. *Total influence* is the influence weight per publication multiplied by the number of publications.

Geller (1978) showed that total influence is the stationary distribution of a Markov chain. This would later become known as undamped PageRank/Eigenfactor, while influence per publication, referred to by some authors as the 'invariant method' (Palacios-Huerta and Volij, 2004) is equivalent to Article Influence (Waltman and van Eck, 2010).

Influence weight, or total influence per outgoing reference, seems to have been paid relatively little attention; Vigna (2016) even called it 'bizarre'. It has been rediscovered several times: Negahban et al. (2012) proposed 'Rank Centrality', which is effectively influence weight applied to ratio matrices. Maystre and Grossglauser (2015) devised an algorithm for k-way comparisons called 'Luce Spectral Ranking', which is identical to influence weight in the $k = 2$ case.

To give the 'influence weight' measure a more distinctive moniker we will call it the **Scroogefactor**, a less unwieldy name which serves to highlight its role in bibliometrics as an alternative to the impact factor and the Eigenfactor. Moreover, *Scrooge*factor emphasises its key feature: the score penalises journal editors who are generous in allowing citations and rewards those who are miserly. This is something that Article Influence score does not quite handle, as its own creators admit:

> 'As is the case with impact factor scores, review journals will score higher [in Article Influence] because of the large number of citations that individual articles in these journals receive. Thus, it can be important for some applications to compare non-review journals with non-review journals and review journals with review journals.'
>
> — West (2010), page 15

Rather than a dichotomy between 'review' and 'non-review' journals (in the JCR, an article is classed 'review' if it cites more than 100 references), a metric which is influence per reference rather than per article (as shall be shown below) allows a much smoother weighting between journals that give out a lot of citations and those that give out relatively few.

But the Scroogefactor has theoretical as well as practical applications: it can be shown that the Scroogefactor provides a direct link between PageRank and the Bradley–Terry model.

Let A be a diagonal matrix with elements $a_{ii} = \sum_{k=1}^{n} c_{ki}$, the column sums of C (that is, the numbers of references in the bibliogra-

phy of each journal). Then (2.25) can be rewritten as the eigenvector equation

$$\mathbf{w} = A^{-1}C\mathbf{w}, \tag{2.26}$$

where $\mathbf{w} = (w_1, \ldots, w_n)$ is the vector of Scroogefactor scores (influence weights). By comparison, (2.1), with $\alpha = 1$, gives the relation

$$\boldsymbol{\pi} = \widetilde{C}\boldsymbol{\pi} = CA^{-1}\boldsymbol{\pi}, \tag{2.27}$$

where $\boldsymbol{\pi}$ is the undamped PageRank (total influence) vector.

The scaled matrix $A^{-1}C$ (visualised in Figure 2.12) is *similar* to the column-stochastic probability transition matrix $\widetilde{C}$, therefore has largest eigenvalue equal to 1. In other words, influence weight is the leading eigenvector of $A^{-1}C$. Moreover, by matrix similarity, $\mathbf{w} = A^{-1}\boldsymbol{\pi}$, so influence weight is PageRank per reference (Geller, 1978).

Theorem 2.1.

Theorem 2.2. *Suppose $C = DS$ is quasi-symmetric. Let $\mathbf{d} = De$ be the vector corresponding to the diagonal elements of D. Let $A = \operatorname{diag}(e^TC)$ be the diagonal matrix with elements equal to the column-sums of C. Then $\mathbf{d}$ is the leading eigenvector of $A^{-1}C$.*

This theorem implies that, under quasi-symmetry, the Stigler/Bradley–Terry model and Scroogefactor scores are the same. In practice the cross-citation matrix is only approximately quasi-symmetric, so the Stigler-model and Scroogefactor scores are highly correlated but not identical (see Figure 3.4).

Proof. We will show that $\mathbf{d}$ is the eigenvector of $A^{-1}C$ corresponding to the eigenvalue 1, and therefore equal to the vector of influence weights.

$$\begin{aligned} A^{-1}C\mathbf{d} &= A^{-1}DSDe \\ &= A^{-1}D(e^TDS)^T \\ &= DA^{-1}Ae \\ &= De \\ &= \mathbf{d}. \end{aligned}$$

This is the leading eigenvector, because $A^{-1}C$ is *similar* to the stochastic matrix CA^{-1} and similar matrices have the same eigenvalues (Newman, 2010, p. 138). Hence $\mathbf{d} = \mathbf{w}$. □

Under quasi-symmetry, the probability transition matrix for the PageRank Markov chain is

$$\widetilde{C} = CA^{-1} = DSA^{-1} = AA^{-1}DSA^{-1} = AD(A^{-1}SA^{-1}), \tag{2.28}$$

by commutativity of diagonal matrices. The expression inside the brackets is symmetric and the product of two diagonal matrices is a diagonal matrix, so the transition matrix is quasi-symmetric.

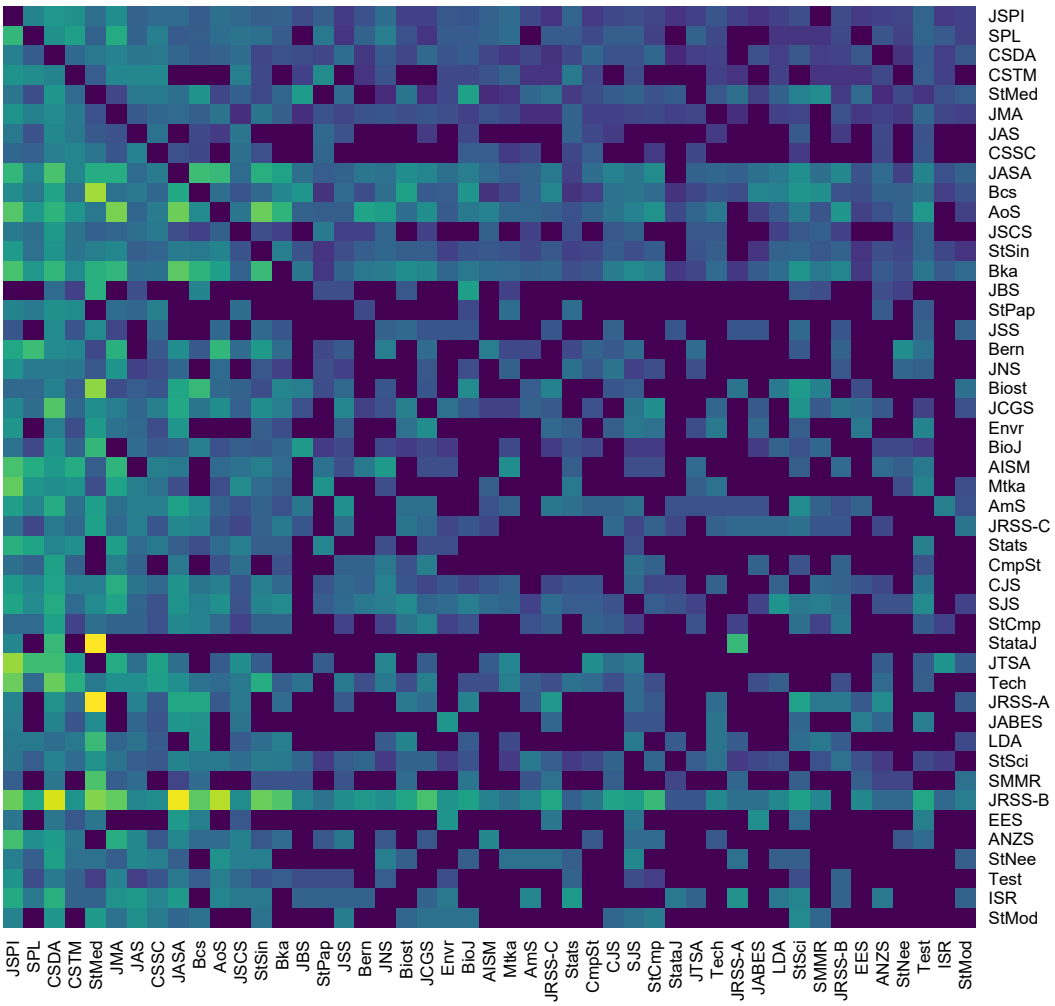


Figure 2.12: A heatmap of the 'Scrooge-adjusted' citation matrix, $A^{-1}C$, where each incoming citation to a journal j is divided by the total number of outgoing citations from j. Journals are sorted in descending order by article count

Thus if Bradley–Terry scores are equal to influence weights, then the corresponding Markov chain is reversible.

The converse is also true: if the Markov chain is reversible, then the transition matrix is quasi-symmetric (Kelly, 1979; McCullagh, 1982), so it can be decomposed in the form $\widetilde{C} = DS$ for some diagonal matrix D and symmetric matrix S. But to transform a transition matrix back to a contingency table is a simple scaling by a diagonal matrix, say A, so

$$C = \widetilde{C}A = DSA = AA^{-1}DSA = A^{-1}D(ASA), \tag{2.29}$$

which is quasi-symmetric, by similar reasoning to above. Thus Bradley–Terry scores are equal to Scroogefactor scores if and only if the network's PageRank Markov chain is reversible.

For our 47 statistical journals, the Stigler-model export scores and Scroogefactor scores are compared in Figure 2.13, and the respective rankings are compared in Figure 2.14. The strength of the correlation between the two metrics, applied to this real data set, is striking.

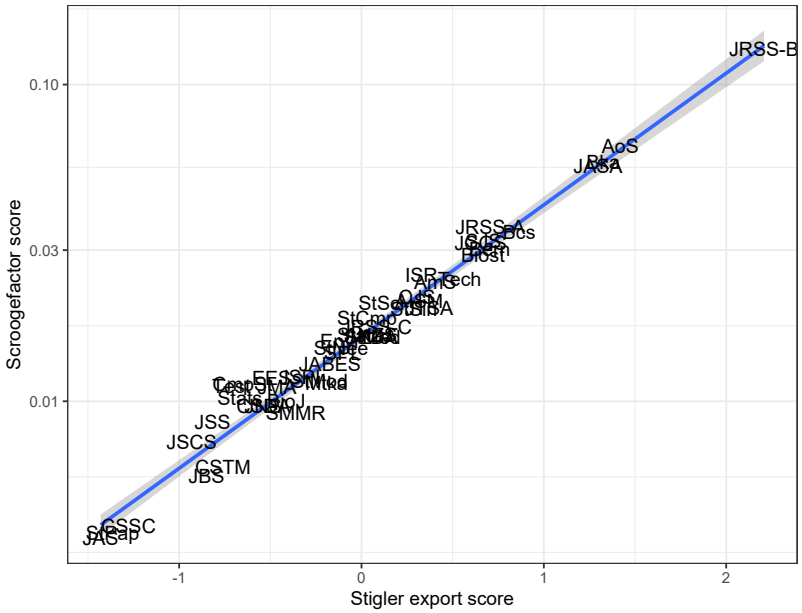


Figure 2.13: Scatter plot of 'Scroogefactor' score (on a log scale) against estimated Stigler export score, with a line of best fit

2.5.3 *Asymptotic efficiency*

For certain special structures, it can be shown via the delta method that the (log) Scroogefactor is an asymptotically efficient estimator for Bradley–Terry scores. Consider a 'round robin' network (tournament) in which every entity cites (beats) every other entity an equal number of times, k. Then the citation matrix is $\mathbf{C} = k\mathbf{e}\mathbf{e}^T$. The corresponding column-stochastic probability transition matrix is then $\mathbf{P} = \mathbf{C}\mathbf{D}^{-1} = \frac{1}{n}\mathbf{e}\mathbf{e}^T$.

We model perturbations of this arrangement by $\mathbf{C}_t = \mathbf{C} + t\mathbf{F}_{ij}$, where t is a parameter for the magnitude of perturbation and $\mathbf{F}_{ij}$ is

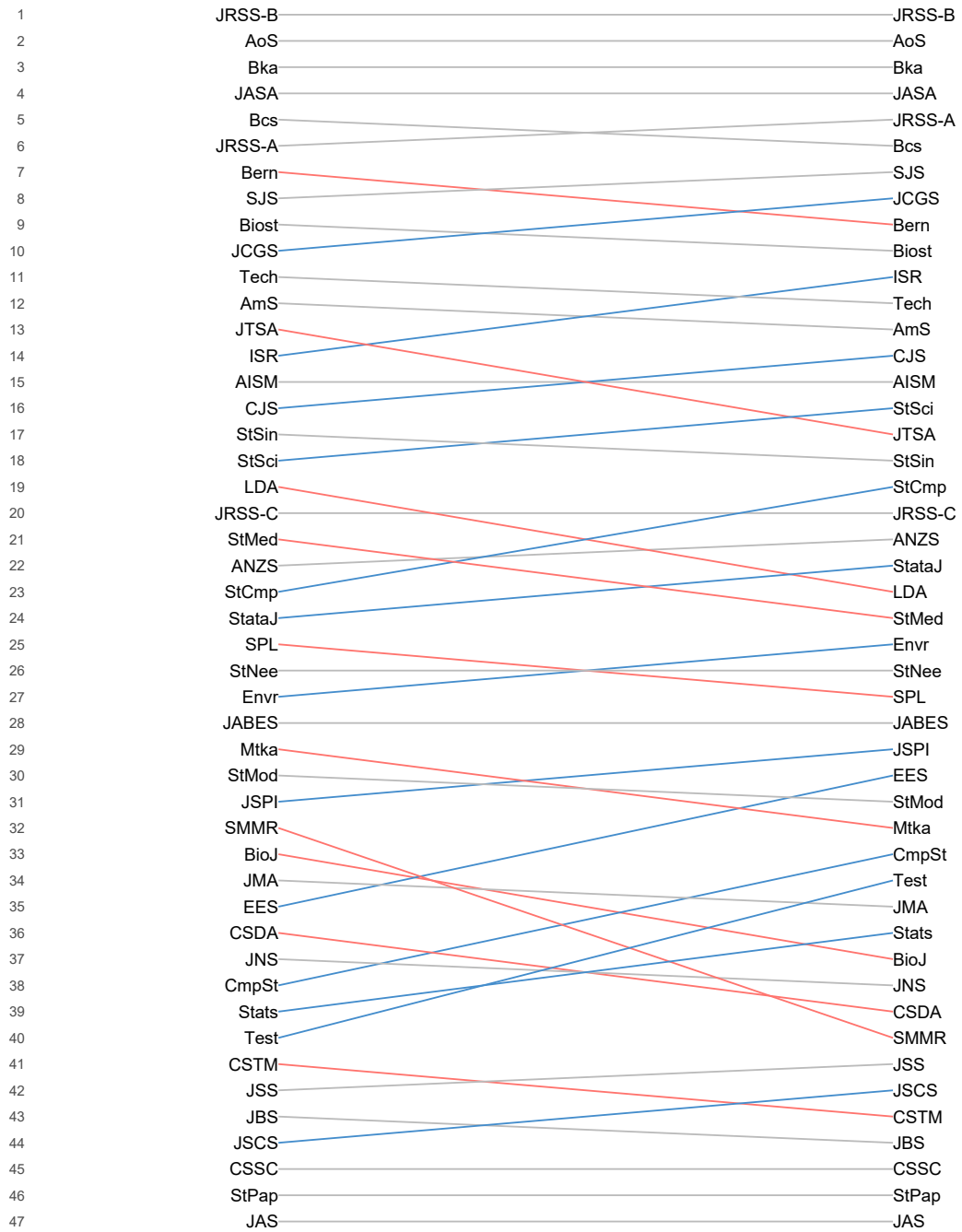


Figure 2.14: Comparison of journal rankings by Stigler-model export score and by 'Scroogefactor' score. Red lines denote journals ranked more highly by the Stigler model, blue lines denote journals ranked more highly by Scroogefactor and grey lines denote journals ranked equally (± 1 place) by the two methods

the $n \times n$ matrix with element (i, j) equal to 1, element (j, i) equal to -1 and all other elements equal to zero. When $t = 0$ then $\mathbf{C}_t = \mathbf{C}_0 = \mathbf{C}$, so the citation matrix is unperturbed.

The derivative of $\mathbf{P}$ with respect to t, for a perturbation of element (i, j), is

$$\frac{\partial \mathbf{P}}{\partial t} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \underbrace{\frac{1}{n^2}\mathbf{e} - \frac{1}{n}\mathbf{e}_j}_{\text{column } i} & \cdots & \underbrace{-\frac{1}{n^2}\mathbf{e} + \frac{1}{n}\mathbf{e}_i}_{\text{column } j} & \cdots & \mathbf{0} \end{bmatrix},$$

where $\mathbf{0}$ is an n-vector of zeros and $\mathbf{e}_i$ is the n-vector whose ith component is 1 with all other components zero. In other words, the ith column of $\partial \mathbf{P}/\partial t$ is all $\frac{1}{n^2}$, except element (j, i), which is equal to $-\frac{n-1}{n^2}$. The jth column is all $-\frac{1}{n^2}$, except element (i, j), which is equal to $\frac{n-1}{n^2}$. Every other column is filled with zeros.

The partial derivative of an eigenvector is (Golub and Meyer, 1986)

$$\dot{\boldsymbol{\pi}}_0 = \boldsymbol{\pi}_0 \dot{\mathbf{P}}^T (\mathbf{I} - \mathbf{P})^\dagger,$$

where a dot denotes partial differentiation with respect to t and † denotes the Moore–Penrose pseudoinverse.

Putting it all together,

$$\dot{\boldsymbol{\pi}}_0 = \begin{bmatrix} 0 & \cdots & \underbrace{\frac{1}{kn^2}}_{\text{element } i} & \cdots & \underbrace{\frac{-1}{kn^2}}_{\text{element } j} & \cdots & 0 \end{bmatrix}^T.$$

By the product rule, the derivative of the unnormalized Scroogefactor is $\dot{\mathrm{SF}}_{\mathrm{un}} = \dot{\mathbf{D}}^{-1}\boldsymbol{\pi} + \mathbf{D}^{-1}\dot{\boldsymbol{\pi}}$. With normalization, $\mathrm{SF}_{\mathrm{norm}} = \frac{\mathbf{D}^{-1}\boldsymbol{\pi}}{\mathbf{e}^T\mathbf{D}^{-1}\boldsymbol{\pi}}$. By the quotient rule

$$\dot{\mathrm{SF}}_{\mathrm{norm}} = \frac{\psi \dot{\mathrm{SF}}_{\mathrm{un}} - \mathrm{SF}_{\mathrm{un}}\,\dot{\psi}}{\psi^2},$$

where $\psi = \mathbf{e}^T \mathrm{SF}_{\mathrm{un}} = \frac{1}{kn}$ denotes the sum of the unnormalized Scroogefactor vector. (Its derivative is zero.)

This yields

$$\dot{\mathrm{SF}}_{\mathrm{norm}}(0) = \begin{bmatrix} 0 & \cdots & \underbrace{\frac{2}{kn^2}}_{\text{element } i} & \cdots & \underbrace{\frac{-2}{kn^2}}_{\text{element } j} & \cdots & 0 \end{bmatrix}^T.$$

The derivative of the log-Scroogefactor, by the chain rule, is

$$\begin{aligned} \frac{\partial \log \mathrm{SF}}{\partial t}(0) &= \frac{\dot{\mathrm{SF}}_{\mathrm{norm}}}{\mathrm{SF}_{\mathrm{norm}}} = n\dot{\mathrm{SF}}_{\mathrm{norm}} \\ &= \begin{bmatrix} 0 & \cdots & \underbrace{\frac{2}{kn}}_{\text{element } i} & \cdots & \underbrace{\frac{-2}{kn}}_{\text{element } j} & \cdots & 0 \end{bmatrix}^T. \end{aligned}$$

Extending this scenario from scalar perturbations, t, to every possible combination of perturbations of the upper triangle [3] of $\mathbf{C}$, we introduce the $\binom{n}{2}$-length perturbation vector $\mathbf{t}$ and calculate

[3] N.B. any perturbation of the lower triangle of $\mathbf{C}$ is equivalent to a perturbation to the upper triangle with opposite sign.

the Jacobian with respect to the same. The Jacobian of log SF is an $n \times \binom{n}{2}$ matrix

$$\mathbf{J} = \begin{bmatrix} + & + & + & 0 & 0 & - & \cdots \\ - & 0 & 0 & + & + & + & \cdots \\ 0 & - & 0 & - & 0 & 0 & \cdots \\ 0 & 0 & - & 0 & - & 0 & \cdots \end{bmatrix},$$

where '+' and '−' represent the positive and negative elements of the partial derivative $\frac{\partial \log \mathrm{SF}}{\partial t}$ (that is, $\pm\frac{2}{nk}$) and where each column corresponds to a perturbation of the upper triangle of $\mathbf{C}$.

If we assume the data are generated from independent binomials (with $2k$ trials for each pairing and success probability $1/2$), the covariance matrix $\mathbf{\Sigma}$ is an $\binom{n}{2} \times \binom{n}{2}$ diagonal matrix with every diagonal element equal to $2k \times \frac{1}{2} \times \frac{1}{2} = \frac{k}{2}$.

Applying the delta method, the resulting first-order approximate covariance matrix for the log-Scroogefactor is $\mathbf{\Sigma}_{\mathrm{SF}} = \mathbf{J}\mathbf{\Sigma}\mathbf{J}^T$, with elements

$$[\mathbf{\Sigma}_{\mathrm{SF}}]_{ij} = \begin{cases} \dfrac{2(n-1)}{kn^2} & i = j, \\ \dfrac{-2}{kn^2} & i \neq j, \end{cases}$$

for $i, j = 1, 2, \ldots, n$.

This is exactly equal to the asymptotic covariance matrix of log-ability scores from a Bradley–Terry (Stigler) model fitted to the same dataset. Hence for an equal-abilities, round-robin tournament, the log-Scroogefactor is an asymptotically efficient estimator for the Bradley–Terry model.

Now consider a different tournament structure where players hold hands in a circle. The corresponding citation matrix has a *cycle* or *circumplex* structure: that is, a band with non-zero entries on the sub-diagonal and super-diagonal and in the top-right and bottom-left corners. This might also be described as a *circulant* matrix, generated by the vector

$$\mathbf{c} = \begin{pmatrix} 0 & k & 0 & 0 & \cdots & 0 & k \end{pmatrix}^T$$

in the first column.

The citation matrix for an n-player circular tournament, where every player cites each neighbour k times, would look like

$$\mathbf{C} = \begin{bmatrix} & k & & & & & k \\ k & & k & & & & \\ & k & & \ddots & & & \\ & & \ddots & & k & & \\ & & & k & & k & \\ k & & & & k & & \end{bmatrix}.$$

The corresponding probability transition matrix is

$$\mathbf{P} = \begin{bmatrix} & 1/2 & & & & 1/2 \\ 1/2 & & 1/2 & & & \\ & 1/2 & & \ddots & & \\ & & \ddots & & 1/2 & \\ & & & 1/2 & & 1/2 \\ 1/2 & & & & 1/2 & \end{bmatrix}$$

and clearly the unperturbed PageRank and influence weight vectors are both equal to $\frac{1}{n}\mathbf{e}$.

Using the same approach as for the round robin tournament, we find the variances of log influence weights (i.e. diagonal entries of the covariance matrix) for the circular tournament are equal to

$$\frac{n^2-1}{6kn}$$

The next covariance terms (i.e. the super/sub-diagonal entries) are

$$\frac{(n-1)(n-5)}{6kn},$$

followed by (in the third band, assuming $n \geq 3$)

$$\frac{n^2-12n+23}{6kn}$$

and so on. These are exactly equal to the asymptotic covariances for a Bradley–Terry model fit to the same data. So efficiency holds for circular tournaments as well as round robin ones.

2.5.4 *Damping and pseudocounts*

So far we have shown that under certain conditions an *undamped* PageRank vector can be scaled to a Scroogefactor score, which can be used as an estimator for the Bradley–Terry model. But what if we are given a *damped* PageRank vector?

The damping transformation (2.1) might be considered analogous to adding *pseudocounts* to a citation matrix to tweak the estimated ability scores in a Bradley–Terry model.

From the previous subsection we have seen that an undamped PageRank, $\boldsymbol{\pi}$, is the leading eigenvector of CA^{-1} and that Scroogefactor, $A^{-1}\boldsymbol{\pi}$ yields a vector equal to the ability scores of a Bradley–Terry model fitted to $C = (CA^{-1})A$. It seems reasonable, then, that damped Scroogefactor should be equivalent to the Bradley–Terry scores of

$$\begin{aligned} C_{\text{pseudo}} &= \left[\alpha C A^{-1} + \frac{1-\alpha}{n} ee^T\right] A \\ &= \alpha C + \frac{1-\alpha}{n} ee^T A \\ &= \alpha C + \frac{1-\alpha}{n} ee^T C. \end{aligned} \tag{2.30}$$

Unfortunately, damped Scroogefactors and Bradley–Terry scores for pseudocounts of the form (2.30) are close but not identical[4]. In fact, the Markov chain for damped PageRank ($\alpha < 1$) is in general not reversible (Gleich, 2015) therefore there does not exist a matrix of pseudo-counts for which Bradley–Terry scores and 'damped' Scroogefactor scores will be exactly equal.

[4] Although there is one nice property: for a given α, the *undamped* Scroogefactor scores of C_{pseudo} are equal to the *damped* Scroogefactor scores of C.

Nonetheless we might like to put a upper bound on the disagreement between damped PageRank and pseudo-Bradley–Terry. Nielsen and Weber (2015) propose a method to find the 'nearest' reversible Markov chain to an irreversible one, where 'nearest' corresponds to minimum distance as measured by the Frobenius norm. In effect, the Bradley–Terry model is finding an alternative 'nearest reversible chain' via maximum likelihood, implicitly measuring distance by the Kullback–Leibler divergence.

Avrachenkov et al. (2010) proposed a different way of damping the PageRank random walk that preserves reversibility, although this changes the problem.

An intuitive method to 'connect' an otherwise disconnected network would be to introduce pseudodata in the form of an extra 'player zero' (or 'journal zero'), who wins and loses against every other player in the network in such a way that quasi-symmetry is preserved. Future work could involve comparing these approaches and finding an appropriate bound.

2.6 *Conclusions*

The Stigler model and the Eigenfactor algorithm are both capable of identifying the top statistical journals, based on one year of data from Journal Citation Reports.

Though the Eigenfactor score increases with journal article count, the Article Influence score corrects for this and provides a ranking correlated with Stigler-model export scores. Based on a statistical framework, the Stigler model has a range of useful properties including size invariance and easy construction of comparison intervals allowing straightforward inferences about generated rankings.

The Eigenfactor Metrics are more widely known but it is a non-trivial task to estimate standard errors for Eigenfactor or Article Influence scores. As such, it is difficult to make inferences about journals with scores that are close together, or to quantify the uncertainty associated with league tables that the metrics produce. In addition, some explicit steps in the Eigenfactor algorithm are difficult to justify, such as the choice of the damping factor or the decision to omit self-citation data.

Through theoretical work in the last section, it was demonstrated that the apparently loosely-related Eigenfactor and Stigler model ranking methods are in fact closely linked and that for quasi-symmetric citation matrices, an influence weight ('Scroogefactor') eigenvector algorithm produces results identical to Stigler-model export scores. Where the quasi-symmetry model fits approximately,

influence weights and export scores are very highly correlated; the former can be computed iteratively to estimate the latter.

3
Inter-field citation modelling

Much of Chapter 2 is centred on the analysis of 18,786 citations between 47 statistical journals from 2001 to 2010. Those data were collected by Varin et al. (2016) from Thomson Reuters' Journal Citation Reports. This chapter will look at a different data set with broader scope.

To investigate patterns of citation behaviour more generally, especially between fields, it is important to analyse data from a range of different disciplines, and on a larger scale where networks may not always be as well-connected. Moreover, whereas the final set of 47 statistical journals was threshed out manually from some 110 in the statistics and probability category, this ad hoc approach to defining fields is not particularly scalable or reproducible.

In this chapter we will model the flow of citation between fields, by aggregating the publications in each field into 'super-journals' and applying the Stigler model and Scroogefactor journal ranking metrics.

We also compare the influence that individual journals have within their discipline, versus the influence they exert on the community as a whole. In some highly specialized sub-fields, localised assessments of the 'top' journals may overlook the impact that certain publications have outside the field. By comparing intra-field influence with wider influence, we can see which publications are more specialized and which play a more interdisciplinary role.

3.1 Field classification

Defining fields is a problem. The definition of a field changes over time, as shown effectively by Moritz Stefaner (2009) in his visualisation of 'The emergence of neuroscience', showing (from clustering based on citation behaviour) that neuroscientific journals emerged from a disparate collection across medicine, molecular biology and neurology to become a well-defined field of their own.

Community detection in networks is a fairly well-studied topic. We will not review the topic in detail here; we explore the relevant literature in more depth in Chapter 4.

Even assuming fields are well-defined and known, there are several ways to tackle global or inter-field citation ranking. One

of the criticisms of the impact factor is that it varies significantly between fields (Seglen, 1997; Amin and Mabe, 2003). Eigenfactor is calculated as a global ranking across fields but faces a journal size bias, while Article Influence tries to control for this but does not distinguish between types of articles, so could still favour review journals (West, 2010).

Source-normalized impact per paper (SNIP) is a metric published by Scopus (Moed, 2010), which—given some definition of the field that the journal covers—weights citation counts per paper according to the different received citation rates between fields. This method goes some way to correcting some of the biases of the impact factor, but was criticised by Leydesdorff and Bornmann (2011) as not a 'proper statistic' as it conflates medians and means, making it unsuitable for standard statistical tests. Leydesdorff et al. (2013) proposed an alternative method for 'field-normalized impact factors' though as a form of impact factor still does not alleviate all of problems the impact factor has *within* fields or sub-fields.

For the following analysis, we rely on Thomson Reuters' classification of journals into fields and sub-fields and assume they are correct. This is not ideal but is not an unreasonable place to start. Later, we will consider alternative ways of grouping the journals.

3.2 *The data*

We have access to a large data set from Thomson Reuters (now Clarivate Analytics). This provides the opportunity to evaluate how journal ranking metrics work on entire academic disciplines. By aggregating all publications in each field into a single 'super-journal', it is possible to model the exchange of citations between disciplines.

Lists of journals in ten fields—biology, chemistry, computer science, engineering, medicine, mathematics, multidisciplinary sciences, psychology and statistics & probability—were collected from the InCites Journal Citation Reports[1] (JCR) database. In the JCR, journals are not neatly assigned to these fields; the ten listed here were amalgamated from some 116 sub-fields.

[1] https://jcr.clarivate.com

The table of journal and field names was merged with a very large data set of citation counts kindly provided by Thomson Reuters. The latter data describe the frequency of citations indexed by the Web of Science, from journals published in 2012 to journals published in 2003–2012. After filtering out those journals not belonging to at least one of the ten fields, 20,146,725 citations between 7,386 journals remain.

Disciplines are not disjoint: 1,116 journals are categorised as belonging to multiple fields. Of these, 980 journals are in two different fields, 124 belong to three and 12 are in four. For example, *Biostatistics* is classed as both biology and statistics, *Statistics in Medicine* is categorised as biology, medicine and statistics and the *Journal of Chemometrics* counts as chemistry, computer science, mathematics

	bio	chem	comp	eng	math	med	multi	phys	psych	stats
bio	1851315	460086	6061	67624	4007	614614	344154	53267	11563	3259
chem	458592	2511534	8853	176202	3093	237368	140925	445399	2647	953
comp	5347	7444	134043	59740	19138	7914	5027	10942	3056	2755
eng	55021	165324	67075	729425	25830	47336	17469	97438	1715	1766
math	2389	2380	15793	21187	194914	1401	3886	21170	2111	5039
med	643604	298207	10560	73582	1797	5404594	375543	37571	142700	2199
multi	326752	179923	4629	18643	3136	236872	139255	111172	17965	603
phys	46304	430906	12965	110681	25778	26985	54521	1726524	415	1042
psych	8195	1430	4680	2906	722	132100	22058	388	320907	553
stats	4524	1083	4096	3200	5645	4671	1523	1336	867	24819

Table 3.1: Citations between fields in 2003–2013 (from columns to rows), rounded to the nearest integer

and statistics[2]. Slightly perversely, only 10 journals in the field of multidisciplinary sciences are also in other fields; *Nature* and *Science* are not among them.

[2] Perhaps unsurprisingly, Thomson Reuters have classified *Biometrika* as a biology and statistics journal. This may have been the case when it was established as a biometrics journal in 1901, but today the publishers describe it as 'primarily a journal of statistics'.

To avoid double counting, citations are fractionally weighted according to the number of possible inter-field interactions. For instance, each citation from *Biostatistics* to *Statistics in Medicine* might reasonably be counted as any or all of six combinations leading from biology or statistics to biology, medicine or statistics. In this case each citation is counted for every one of these pairs with weight 1/6, thus preserving the count of one citation overall. An alternative method of counting would be to discount any citations between sets of fields that overlap; an analysis based on this approach gives similar results to the following analysis.

3.3 *Visualisation*

Table 3.1 shows the resulting weighted citation counts for the ten academic disciplines. A table is not, however, the most effective way[3] to visualise the flow of citations between the fields. As a static visualisation, a conventional node-link diagram may not be the way to go either, because we have a complete tournament (there are no zeroes in the Table 3.1) with very many weighted, directed edges. In the analysis of the 47 statistics journals we considered the use of heatmaps, with a mosaic of pixels coloured according to the relative number of citations in each cell of the table. In this analysis there are relatively few (super)-journals so there is an opportunity to try a different approach.

[3] In my opinion.

Figure 3.1 shows a chord diagram, which represents the citation counts radially with their flows drawn as quadratic Bézier curves (Gu et al., 2014). It is clear to see that much of the network is dominated by intradisciplinary citations. In medicine, citations within the field account for more than all other citations, sent or received. Biology is about as likely to cite other fields as to cite itself. Multidisciplinary sciences journals are more likely to cite other fields. Disciplines that might be considered more 'theoretical' (Figure 3.3)

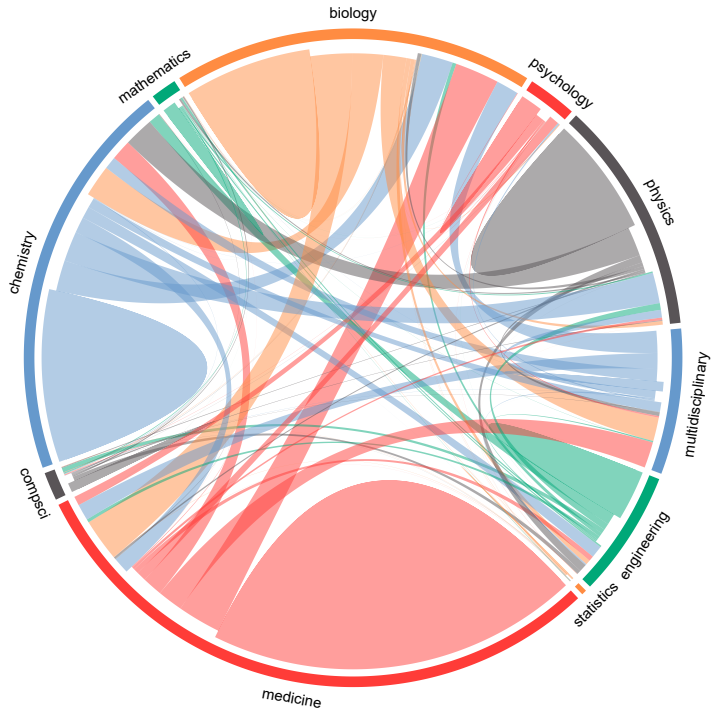


Figure 3.1: Chord diagram of the flow of citations between academic fields. Citation arcs are inset from and the same colour as their field of origin. The overall width of arcs that do not lead anywhere ('mountains') represent field self-citations.

are much smaller fields than the lab sciences in terms of the overall numbers of citations issued and received.

Figure 3.2 shows the same kind of visualisation as Figure 3.1 but with field self-citations omitted, representing a net 7,109,395 non-self-citations. This makes it easier to see the interdisciplinary relationships. For example, medicine and biology exchange a large volume of citations, as do the three 'core' sciences of biology, chemistry and physics. Engineering and computer science seem to be closely linked. Psychology depends heavily on medicine for external citations[4].

[4] Within this network, at least. Had more social sciences fields, such as economics, been included, the picture might be different.

3.4 *Field rankings*

Using the 10×10 inter-field citation matrix, we can apply the Stigler model and the Scroogefactor algorithm (see Section 2.2) to obtain ability scores for the academic disciplines.

Stigler-model export scores and Scroogefactor scores were estimated using the `scrooge` package[5]. It was not possible to compute Eigenfactor or Article Influence scores because the downloaded field-level data did not include article counts (but numbers of 'Citable Items' are available in the JCR database and could be accessed in a future analysis).

[5] `https://github.com/Selbosh/scrooge`

The fields' Scroogefactor scores against estimated Stigler-model export scores are visualised in Figure 3.4. Both scores yield the same ranking. The scores themselves are very highly correlated,

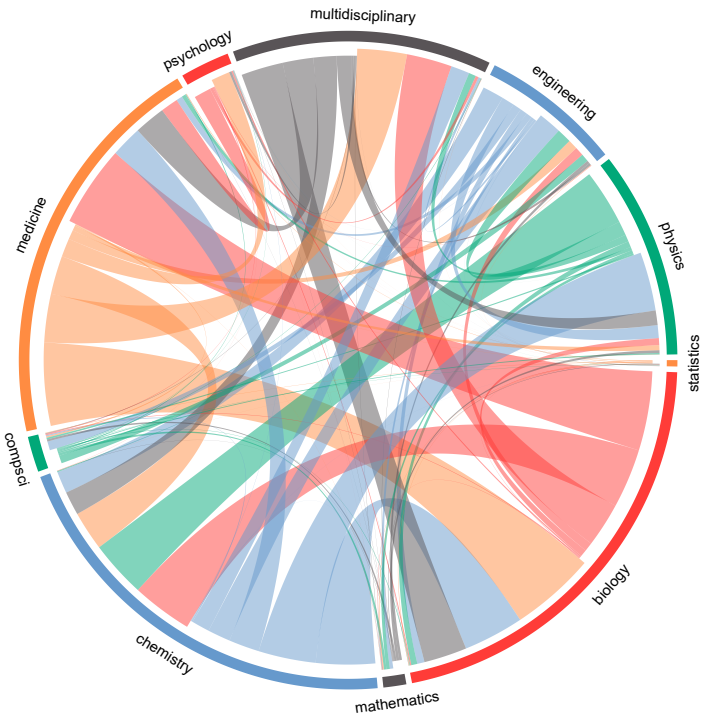


Figure 3.2: Chord diagram of the flow of citations between academic fields. Citation arcs are inset from and the same colour as their field of origin. Field self-citations have been omitted.

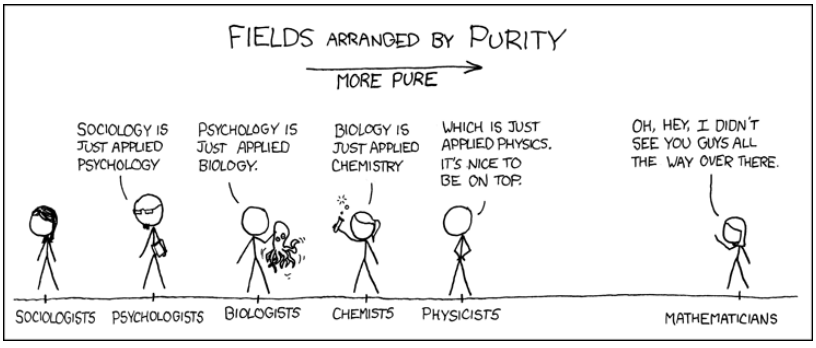


Figure 3.3: Fields arranged by purity (Monroe, 2008)

with Pearson correlation coefficient 0.9981.

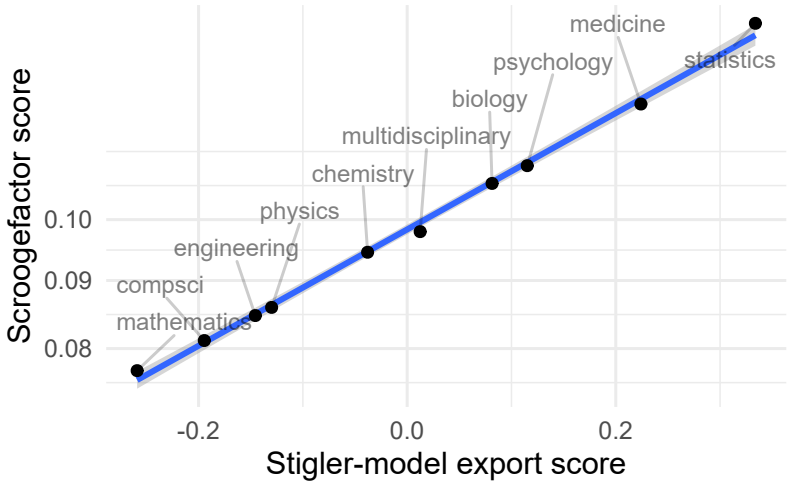


Figure 3.4: Scatter plot of Scrooge-factor scores (on a log-scale) against estimated Stigler model export scores, with a line of best fit

With the exception of statistics & probability, which leads both rankings, the order of academic disciplines seems to follow that of Figure 3.3, with more 'pure' fields at the bottom and more 'applied' subjects at the top. This might seem counter-intuitive—surely the flow of 'influence' should be from theory towards applications? Or it makes sense: applications highlight problems to motivate theoretical and methodological research. It is important to remember that these data only represent citations from journal articles published in 2012 to other journal articles published in the preceding ten years.

The fitted 'quasi-Stigler' model (Section 2.2) includes an extreme level of overdispersion, with $\phi = 1062.35$. The large amount of overdispersion could imply a lack of fit, which may be assessed via the analysis of *journal residuals*, defined in Section 2.4.3. In this case we actually have *field residuals*, though the definition is the same. They should, under the Stigler model, be approximately normally distributed and uncorrelated with the export scores. Figures 3.5 and 3.6 provide field residual plots for the fitted Stigler model.

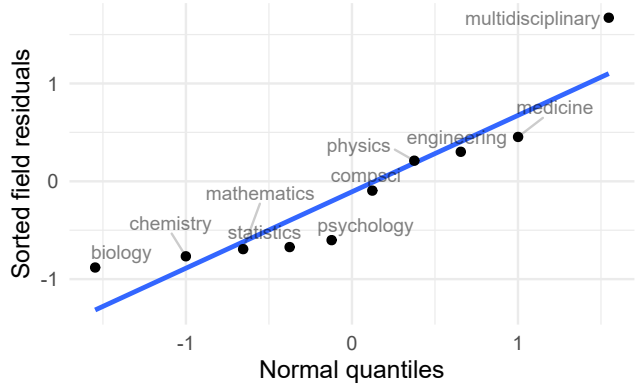


Figure 3.5: Normal Q-Q plot of field residuals for the fitted Stigler model

For the most part, the field residuals appear to be normally distributed and uncorrelated with the fitted export scores. However,

multidisciplinary sciences appears to be an outlier. The field residual for this discipline is large and positive, implying that it performs systematically better than predicted by the model against strong opponents (Varin et al., 2016). In other words, multidisciplinary science journals receive more citations than expected from highly-ranked fields.

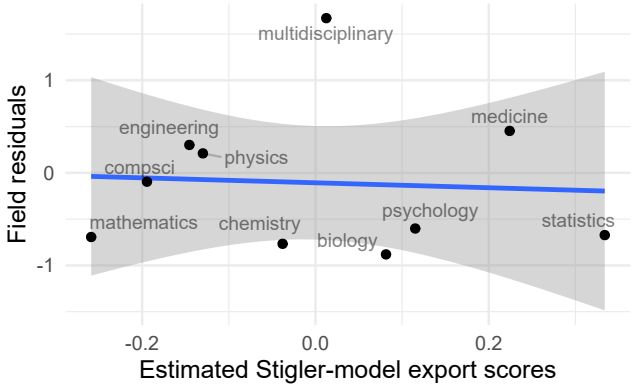


Figure 3.6: Field residuals against export scores for the fitted Stigler model

This provides some evidence against the existence of a simple one-dimensional scale on which every field can be given an unambiguous ranking. It appears that the field of multidisciplinary sciences somehow wants to be ranked more highly than medicine but lower than chemistry and physics, which is not possible in this tournament.

If we remove multidisciplinary sciences as an outlier and model the remaining fields, then the field residuals appear normally distributed with constant variance. However, the estimated parameter of dispersion for the corresponding quasi-Stigler model is $\hat{\phi} = 274.8$, which still seems very large.

To quantify estimation uncertainty in the Stigler model, we use the methods of *quasi-variances* described in the previous chapter. A centipede plot of the Stigler-model export scores—with comparison intervals based on their quasi-standard-errors (the square root of the quasi-variances), computed using the `qvcalc` package (Firth, 2015)—is given in Figure 3.7.

The high overdispersion produces rather large variance estimates. The wide, overlapping 95% comparison intervals imply that many fields are not significantly differently ranked from other fields, suggesting that these disciplines are no more likely to cite or be cited by another in the network, which seems implausible. We do observe, however, a distinct hierarchy for medicine, biology, chemistry and mathematics, in reverse order of 'purity' (Figure 3.3).

Exactness of a quasi-variance approximation may be summarised by the relative errors of quasi-standard-errors from their corresponding standard errors derived from the full variance-covariance matrix. Firth and de Menezes (2004) suggested reporting the 'worst' relative errors and the `qvcalc` package includes these in the summary output. The distribution of relative errors, not just their min-

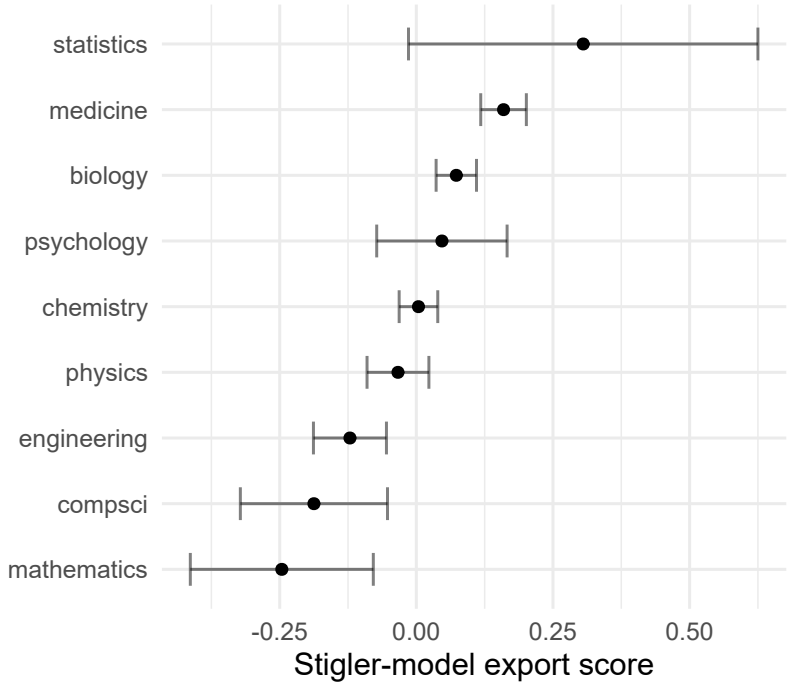


Figure 3.7: Centipede plot of estimated field export scores and 95% 'comparison intervals' (Firth and de Menezes, 2004) for 2003–2012 JCR data. The points represent estimated field export scores; their error bars correspond to $\pm 1.96 \times$ quasi-standard-error of each score. The field of 'multidisciplinary sciences' has been excluded as an outlier

imum and maximum, may be a better indicator of problems with quality of a quasi-variance approximation. It is visualised in Figure 3.8.

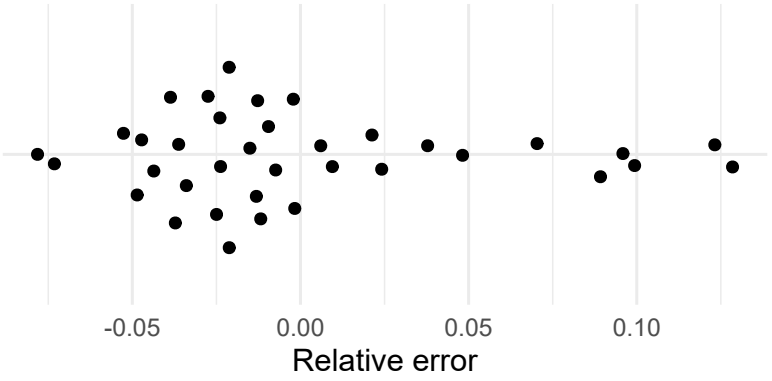


Figure 3.8: Bee swarm plot showing the distribution of relative errors of the quasi-variance approximation to simple contrasts in the fitted Stigler model

The distribution of relative errors is slightly skewed, but it is unimodal and most errors are small[6]. The worst relative errors are −7.8%, between medicine and physics; and +12.8%, between biology and medicine.

The inter-field citation model has index of dissimilarity $\hat{\Delta} = 0.01296$. In other words, about 1.3% of the citations would need to be reassigned for the fitted model to match the observed data exactly. Accuracy of 99% does not seem bad, but compared to what? This number could be employed to compare two competing models, but may not be very useful on its own.

[6] The definition of 'small' is somewhat arbitrary. Errors with magnitude greater than, say, 20% might be considered problematic and a reason to avoid quasi-variances.

3.5 *Resampling*

An alternative way of estimating the variance-covariance matrix of the Stigler model is to use a resampling procedure. The advantage of such an approach, rather than quasi-likelihood estimation, is that it can be applied to non-'statistical' algorithms such as the Scroogefactor as well. Then uncertainty intervals can be compared between models as well as between (super)-journals.

One implementation (not shown here) is to employ a kind of multinomial sampler, which appears to generate comparison intervals much narrower than those computed using quasi-likelihood estimation. Rosvall and Bergstrom (2010) suggested a parametric network resampler that draws independent Poisson distributions with the observed link weights as their means. Mirshahvalad et al. (2013) showed, however, that Poisson citation resampling can underestimate the variance of link weights and that a more sophisticated parametric bootstrap requires article-level data.

To minimise the number of assumptions made and to investigate the phenomenon of overdispersion in our Stigler model, here we adopt a non-parametric jackknife approach (Efron, 1979).

In each replicate, the algorithm draws, without replacement, 90% of the citations from the cross-citation matrix. The citations are stratified by year, so 90% of the citations from each year 2003–2012 are sampled. Then the Stigler model is fitted and the vector of field export scores is returned. The process is repeated independently many times and each of the sample score vectors is recorded. We then estimated a sample variance-covariance matrix for these data, using the correction

$$\mathrm{Var}_{R,m}(\boldsymbol{\mu}) = \frac{N-m}{m}\frac{1}{R}\sum_{r=1}^{R}(\hat{\boldsymbol{\mu}}_r - \boldsymbol{\mu})^2, \tag{3.1}$$

where $R = 10{,}000$ is the number of replicates, $N = 20{,}146{,}725$ is the total number of citations, $m = \frac{N}{10}$ is the number deleted, $\hat{\boldsymbol{\mu}}_r$ is the sample ability scores vector for replicate r and $\boldsymbol{\mu}$ is the scores vector computed from the full citation matrix.

From (3.1), we then apply a quasi-variance approximation (using `qvcalc`) and compute comparison intervals. The resampled comparison intervals for the Stigler model are shown in Figure 3.9.

As we do not rely on quasi-likelihood estimation to compute the variance-covariance matrix, exactly the same method can be applied to the ('non-statistical') Scroogefactor algorithm as well. The resulting resampled Scroogefactor comparison intervals are presented in Figure 3.10.

The similarity between the resampled Stigler-model and Scroogefactor score comparison intervals is striking, though perhaps not surprising, given the strong correlation already seen in Figure 3.4. Moreover, both sets of intervals are far narrower than those from the overdispersed quasi-Stigler model (Figure 3.7).

The worst relative errors in the quasi-variance approximations to

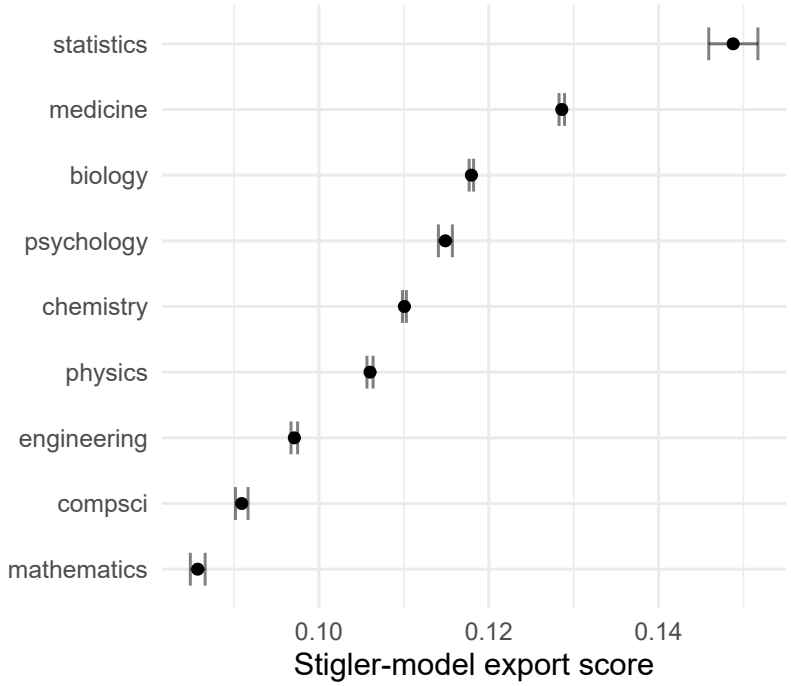


Figure 3.9: Estimated Stigler-model export scores for the nine fields. Error bars are 95% comparison intervals, based on a stratified delete-10% jackknife with 10,000 replicates

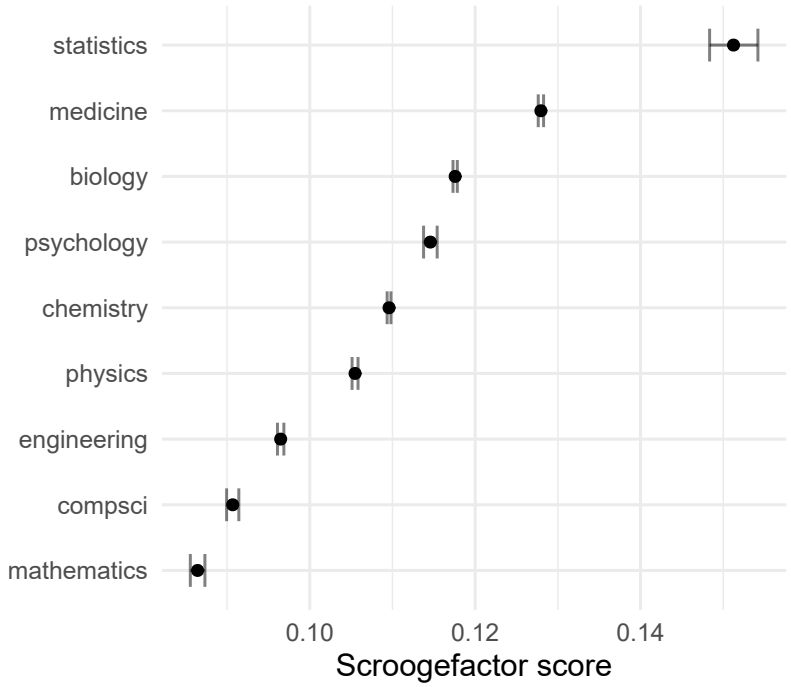


Figure 3.10: Estimated Scroogefactor scores for the nine fields. Error bars are 95% comparison intervals, based on a stratified delete-10% jackknife with 10,000 replicates

the resampled variance-covariance matrices are –7.6% and +12.3% (Stigler model) and –7.7% and +12.4% (Scroogefactor). These are comparable to the relative errors for quasi-variances computed from the quasi-Stigler model.

If these intervals are to be believed, the field ranks are robust: no pair of 95% comparison intervals between fields are overlapping, implying each field's score is significantly different from that of other fields, in both Stigler model and Scroogefactor. The relative magnitudes of variances between models are visually similar, too, and there is a size effect: small fields such as statistics and mathematics have wider comparison intervals than fields with more intense citation flow, such as medicine or biology.

This raises a problem: is there something wrong with quasi-likelihood estimation or with the resampling algorithm as implemented? Why do they give such starkly different results? This was explored at the end of Chapter 2 and may form the basis for future work.

3.6 Modelling at scale

Having demonstrated the techniques for fitting and evaluating inter-field citation models on a set of 10 human-generated categories, we can set our sights on a larger scale analysis. Whereas Web of Science subject categories have been generated from a fairly opaque process, there is nothing to stop us fitting Stigler models to communities that have been generated algorithmically using a reproducible procedure.

We therefore now consider a database of citations—again from Web of Science—but divided automatically into disciplines according to the *Infomap algorithm*, further details of which are provided in Chapter 4. The algorithm, when applied to 29 million citations between 11,000 academic journals in 2006–2015, yields 69 communities of journals. Unlike Web of Science subject categories, these communities do not overlap, so fractional counting of citations is unnecessary.

On inspection, the groupings seem sensible and almost every community can be assigned a human-understable name. For example, most of the statistics journals seen in Chapter 2 are also found in this dataset, nearly all of them inhabiting community '24', along with many more publications with a clear statistical orientation. With little hesitation we can therefore label this community 'statistics'. Most of the other categories are similarly easy to assign names, but we also retain the numeric identifiers in case anybody reproducing these analyses disagrees with our naming scheme and wishes to assign different labels.

For obvious practical reasons, we do not print a list of all the journals and their assigned clusters here, nor is it space efficient to present an exhaustive analysis of all the possible intra-field models on paper. Instead, we invite the reader to explore the data and

results via an interactive visualisation at `https://selbydavid.com/influence/`. The interface (programmed in D3; Bostock et al., 2011) presents an inter-field influence ranking of the 69 communities, as well as the results of multiple intra-field analyses for each one, complete with comparison intervals.

Code to reproduce the analysis is available at `https://selbydavid.com/influence/analysis.html`

Some select results are presented here; please visit the web site to explore other results. The large-scale analysis was assisted by the use of the R package `BradleyTerryScalable` (Kaye and Firth, 2017), which allows fast fitting of Bradley–Terry models through vectorized C++ code (Eddelbuettel and Balamuta, 2017).

3.6.1 *'Journal zero' and the 'other fields' superjournal*

Regularization of the Stigler-model parameter estimates might be achieved with an appropriate prior distribution and performing a Bayesian analysis; however here we introduce a 'player zero', or more appropriately a 'journal zero', which cites and is cited by all other journals equally. Any journals (or superjournals) with a very small number of citations, given or received, are likely to see their scores shrink towards the mean (i.e. 0, in logit space).

For a feeling of familiarity with PageRank, but not for any particularly theoretically-justified reason (which is also similar to PageRank), we will use the hyper-parameter $\alpha = 0.15$, analogous to a damping factor. That is, each 'journal zero' cites and is cited

$$0.15 \times \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}}{2n}$$

times, where c_{ij} is the number of citations from journal i to journal j, and n is the total number of journals in the community (not including the 0^{th} journal, of course).

An advantage of this approach, rather than a full Bayesian model, is the ability to calculate PageRanks or Scroogefactors on the augmented data, since a network with a 'journal zero' in it is just like any other citation network.

It was not possible to use `qvcalc` (Firth, 2015) to compute quasi-variance estimates for the intra-field rankings, as the current version of the package is not optimized for large datasets, and some of our communities—including microbiology, physics/chemistry and general medicine—contain over 1,000 journals each. Instead, we wrote our own implementation[7] in C++ and optimized the squared log distance using a standard BFGS library.

[7] See `https://selbydavid.com/influence/analysis.html#quasi-variances` for the code

When producing within-field rankings, it is easy to lose sight of the wider context. Namely: a journal that is highly regarded by peers within its (possibly highly specialized) field may not be as well-recognized outside the field. When making decisions about interdisciplinary research, or sub-fields that sit within larger ones, it may also be valuable to see whether the wider research community agrees with a field's internal ranking. That is to say, a theoretical statistics journal may be highly regarded by academic statisticians,

but to non-statisticians, software or medical statistics journals might be regarded as more important.

With this in mind, we introduce to each intra-field ranking an entity known as the 'other fields' journal, a super-journal made up of all of the thousands of journals that lie outside the field of interest. As this superjournal is so large, it also has a regularizing effect, negating the need for a 'journal zero' in this case.

What is of most potential interest is how journals' relative ranks change (or not) between the insular intra-field ranking and the wider ranking including the 'other fields' super-journal. On the web site, the transition between these two rankings is animated, which makes clear how slight or significant the differences are: a journal that is equally prominent within its own field and in wider academia may stay in place, whereas one that receives little recognition within its assigned community whilst being cited by other fields may shoot up in the rankings, or drop down the league table if the converse were true. To give a clearer representation of these changes in score, we can also plot the 'intra-field influence' and 'wider influence' scores against one another in a scatter chart. There is not enough space to print all these different graphs here (not least for the communities with > 1000 journals) but we present a selection of representative results in Section 3.6.3.

3.6.2 Inter-field results

The full inter-field ranking of the 69 communities is presented in Figure 3.11 (and is the initial state of the interactive visualisation online). The pattern is not quite as obvious as in the earlier analysis of 10 fields, though it follows a roughly similar trend where applied subjects seem to export more influence than purer fields. For example, the core sciences of physics and chemistry received low rank, along with engineering, whereas medicine or biological science and their various subdisciplines achieve high ranks. Mathematics is below applied mathematics. Social sciences do fairly well.

Statistics is ranked near the top, which is surely the sign of a valid analysis[8]. Psychometrics leads the table, which makes sense as it is an applied field concentrated in a small number of journals. Notably, bibliometrics appears near the bottom of the league, which could be construed as an indictment of the insularity of that field. Meanwhile, it appears that recent particle physics research has little influence on other fields, unless it is published in a radiology journal.

[8] There may be some bias in this assertion.

There are two fields made up of mostly non-English-language publications: Romanian chemistry and Brazilian agronomy. These appear to receive little recognition from the other publications in the Web of Science, which are mostly in the English language. On seeing this latter result, the reader may wonder why the Infomap algorithm has produced one community specifically for agronomy in Brazil, and another for agronomy more generally. Apparently

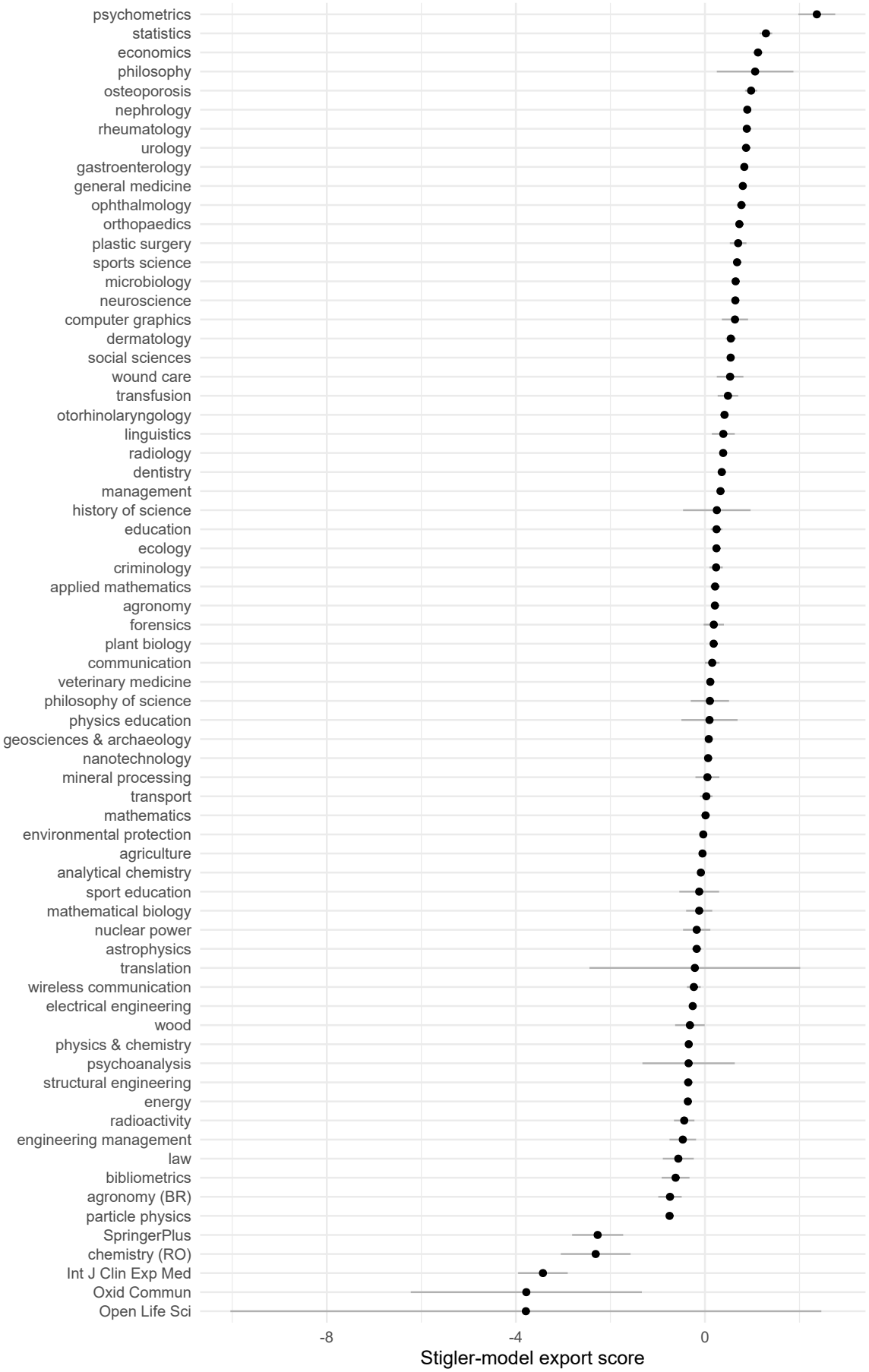


Figure 3.11: The 69 communities in 2006–2015 Web of Science citation data, ranked according to Stigler-model export scores, with 95% comparison intervals

this is not an anomalous result: Dr Jacob van Etten, a researcher based in Costa Rica, helpfully informed us that the Brazilian and South American agronomy field is 'quite insular, partly due to language issues but also the inward looking culture of *Embrapa*, the Brazilian agricultural research agency, with its own journals'. This situation is therefore confirmed in the algorithmic results.

Other outliers include singleton communities such as the journal *SpringerPlus*. Interestingly, the data ended the year before Springer stopped publishing the open-access journal *SpringerPlus* in 2016, so the data appear to confirm what the publisher had apparently learned around the same time: that the journal was failing to gain traction in the scientific community. The remaining singletons mostly seem to be fairly obscure journals published in Central or Eastern Europe.

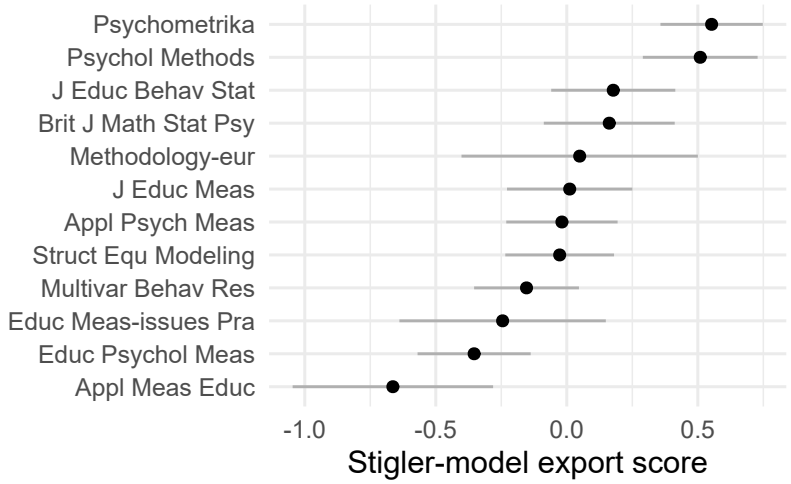
Meta-subjects, such as the history and philosophy of science, are so small that their scores are indistinguishable from zero. Philosophy has a wide comparison interval too, perhaps not because the field is obscure per se, but because philosophers prefer publishing books over journal articles (see Table 6.2).

3.6.3 Intra-field results

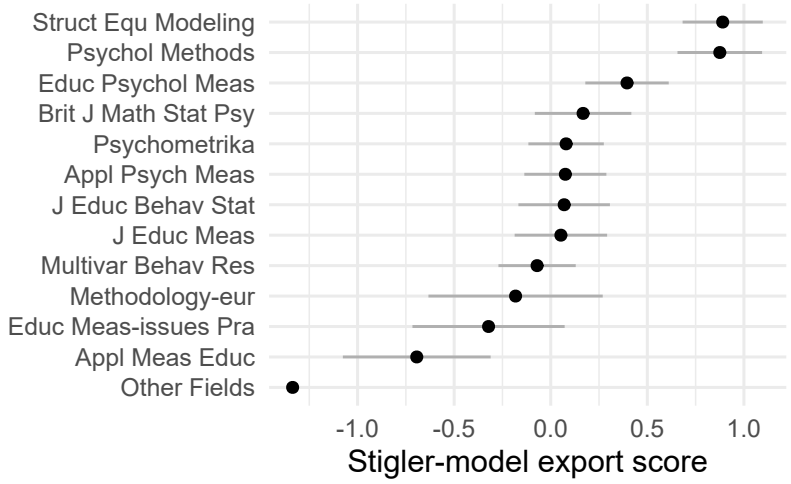
Here we present some results of the rankings *within* fields. Visit the web site to explore more of them, especially large communities, and to see the animated versions of these graphics.

Since psychometrics tops the table, we start there. In Figure 3.12 we can see the within-field ranking of these journals, modelled with and without the influence of citations from other fields. That the 'other fields' super-journal places bottom is to be expected, suggesting that this group of psychometrics journals is quite insular and does not cite other journals very often. Also interesting is the dramatic change in ranking that results. Whereas the journal *Psychometrika* appears (with some uncertainty) to lead the field according to its peers—with *Psychological Methods* a close second—it is *Structural Equation Modeling* that jumps to the top of the ranking when outside citations are included, and *Psychological Methods* stays in second place. This implies that psychometricians may regard the journal *Psychometrika* highly, but *Structural Equation Modeling* perhaps has more recognition among non-specialists, and *Psychological Methods* is equally highly regarded within the discipline and beyond.

The comparison of within-field versus wider influence is made more explicit in Figure 3.13, which shows clearly how the journals *Structural Equation Modeling* and *Educational and Psychological Measurement* may export a great deal more influence outside their field than they do to other specialist psychometrics journals, causing a sudden jump in their relative influence score. Conversely, *Psychometrika*'s score drops when wider influence is taken into account. *Psychological Methods* has a similar estimated score from both



(a) Intra-field influence



(b) Wider influence

Figure 3.12: Stigler-model export scores for journals within the field of psychometrics, with 95% comparison intervals

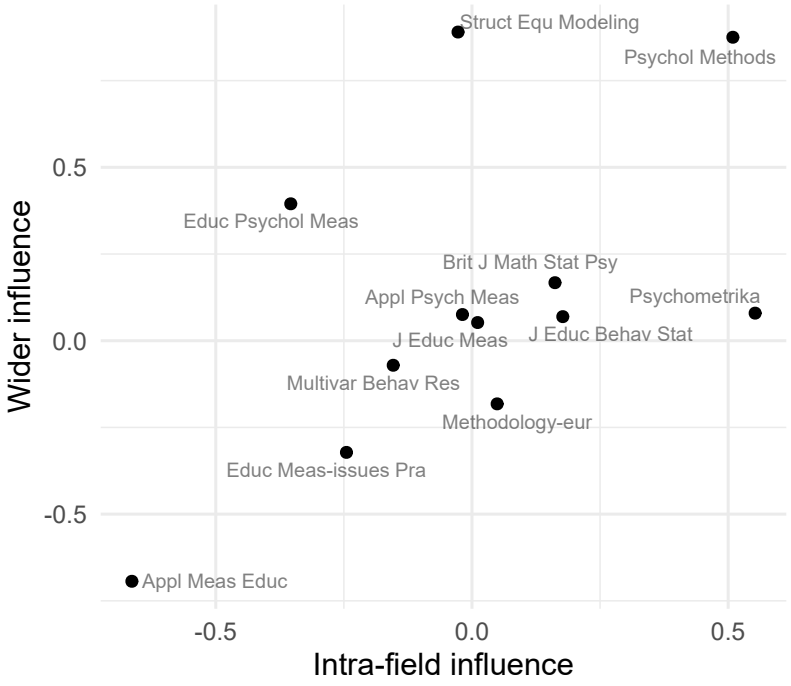


Figure 3.13: Comparison of Stigler-model export scores in the field of psychometrics, with and without the influence of citations from other fields

models, as do most of the other journals in this field.

Now let's consider statistics, which comes second in the inter-field league table, implying that generally it exports more influence than it imports from other disciplines. There are 86 journals in this community. A scatter plot of wider influence versus intra-field influence is given in Figure 3.14. Unlike in psychometrics, the ranking seems fairly robust to the inclusion of external citations. As in previous analyses, the prominent journals *JRSS-B*, *Annals of Statistics*, *JASA* and *Biometrika* are highly ranked. The *Journal of Machine Learning Research* is also in the top five journals by both methods.

Larger discrepancies begin to appear for software journals, namely the *Journal of Statistical Software* and *Stata Journal*, suggesting that these publications are paid more attention by non-specialist statisticians (perhaps in epidemiology or other areas of applied statistics) who publish elsewhere, than by the theoretical or methodological researchers who publish in dedicated statistics journals. An exception to this pattern is the *R Journal*, which has a similar score whether or not external citations are included. We posit that this is because the R language has broad and growing popularity, it is a relatively new journal (launched in 2009), and unlike the older *Journal of Statistical Software* (from 1996) is not nominally limited to only statistical applications.

Finally, we look at mathematics, which comes in the middle of the inter-field league table. This community contains 401 journals, which is too many to list here, but the plots can be explored interactively online. A scatter plot, which shows a remarkably strong correlation between intra-field and wider influence scores, is presented

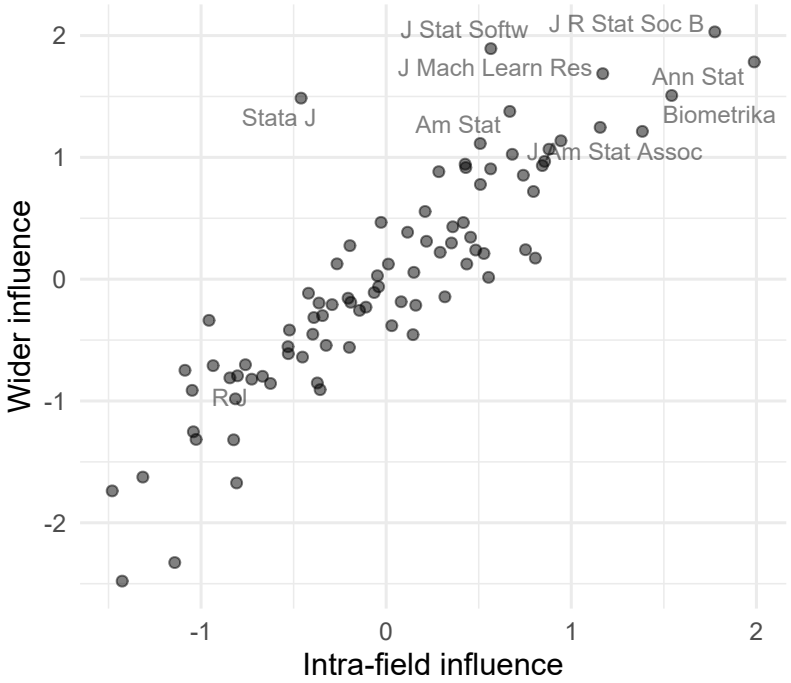


Figure 3.14: Comparison of Stigler-model export scores in the field of statistics, with and without the influence of citations from other fields

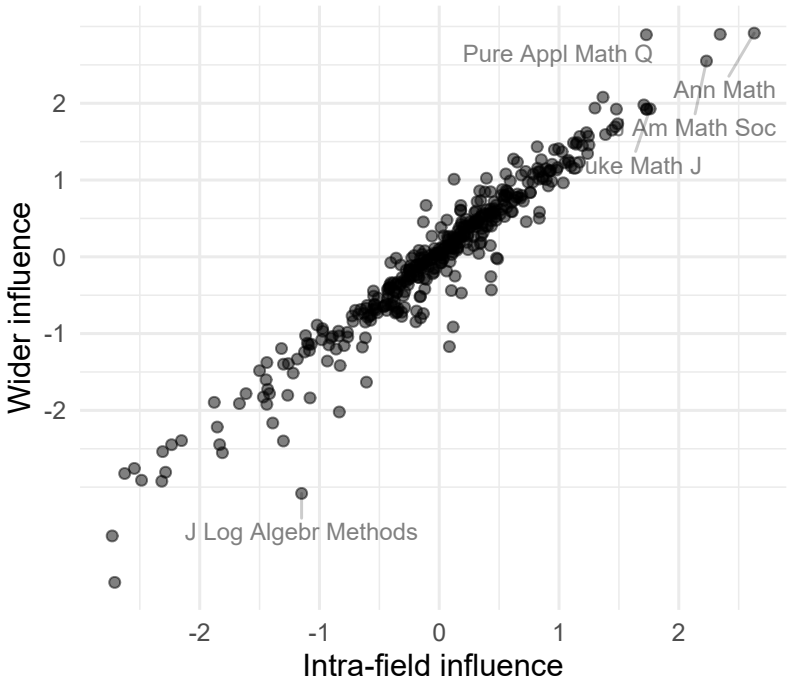


Figure 3.15: Comparison of Stigler-model export scores in the field of mathematics, with and without the influence of citations from other fields

in Figure 3.15, with a handful of prominent journals labelled. We note that the highly-regarded *Annals of Mathematics* and *Acta Mathematica* journals lead the field by both intra-field and wider influence measures. For the most part, mathematical journals have similar influence within their field as they do more widely. Given the field's position in the inter-field ranking, this could simply be because not enough citations are coming from outside mathematics to affect the results noticeably. It may also be a peculiarity of this pure mathematics grouping: all the application-focussed publications (which tend to be most sensitive to the inclusion/omission of external citations) may have simply been placed in other communities, such as those labelled applied mathematics, mathematical biology and statistics.

3.7 Conclusions

In this chapter, we have extended the techniques introduced in Chapter 2 to applications broader in both scale and scope. By aggregating journals into fields we can demonstrate the dynamics of academia as a whole, and gain a better insight into the role of interdisciplinary citation exchange at the disciplinary and journal level.

To our knowledge, direct modelling of interdisciplinary influence is something overlooked in many bibliometric analyses, despite the goal of many metrics and research assessments ostensibly being to measure 'impact'. Paired-comparison models applied to human-curated and algorithmically-generated journal classifications appear to reveal the flow of influence from applied, especially biological subjects to more theoretical or fundamental ones over a ten-year citation window.

The effect of extraneous citations on a community ranking is also considerable, suggesting that neither global rankings, such as the impact factor, nor strictly local ones, such as within-field citation networks, can comprehensively describe researcher behaviour. That journals—and, by extension, individual researchers and their publications—can attain such distinct reputations within, versus between fields, implies that subject-wide performance measures, indiscriminately applied, may overlook influential research.

Conclusions must be tempered by the assumption that journals are a reliable way of aggregating papers on a specific topic or by a particular corps of researchers. Some multidisciplinary magazines such as *Nature* and *Science*, or mega-journals, such as *PLoS One*, are large and cover a range of subjects. Notably, the Web of Science puts such periodicals in a 'multidisciplinary' category rather than assigning them to one or more traditional disciplines. A future analysis might try downweighting or excluding such publications in the data. Alternatively, but perhaps unfeasibly, one could dispense with journal identities and instead cluster individual articles into fields. Ideally this could be achieved via topic models based

on key words, abstracts or full text, then measuring the influence between the resulting groups. Examples of such models for discovering topics from textual data include Griffiths and Steyvers (2004) and Williamson et al. (2010). Here, however, we concentrate on those techniques possible using citation data only. Another way might be to assign papers to groups based on the departmental affiliation of their authors, but this may also be a difficult process to automate.

These present results based only on Web of Science data, so it may also be valuable to attempt to replicate the results using an alternative citation database, such as those discussed in Chapter 5. Another logical extension would be to fit the within- and between-field models for citations over different time windows, to see how intradisciplinary and interdisciplinary influence might change over time.

4
Citation communities

Which film is better, *The Shawshank Redemption* or *The Lion King*? It is not necessarily a reasonable comparison to make—though both pictures are from 1994, they fall in different genres, aimed at different audiences, produced using different media. In the same vein, it may not be productive to compare academic publications from different disciplines using identical standards, because researchers in different disciplines do not all behave in the same way.

Assume, then that either we only wish to compare publications to others in the same field, or that we have some appropriate method for normalising the effect of being from a different field. Both cases depend on the notion of a field or academic discipline being well defined. But is this really the case?

Clarivate Analytics' Web of Science assigns each journal to one or more of 232 categories. For example, *The Lancet* belongs to "general medicine", whilst *Biostatistics* is in both "statistics & probability" and "mathematical & computational biology". For some reason, interdisciplinary journals such as *Science*, *Nature*, *PLoS One* and *PNAS*, instead of belonging to many categories are actually in just one: "multidisciplinary sciences". Some journals are misclassified: the theoretical statistics journal *Biometrika* is considered, presumably based on the name alone, also to be a biology publication.

Leydesdorff et al. (2016) suggested that a good journal classification scheme should be transparent and reproducible by others. It would be hard to make the case that the Web of Science categories pass this test[1], yet they have 'become accepted as "best practice" among bibliometricians' (Leydesdorff et al., 2016). It is acknowledged, however, that "there are no unique or universally valid classifications of journals" (Leydesdorff et al., 2017).

[1] According to Fleck (2013): 'Neither the definition of disciplines, nor the selection of journals for the Web of Science/Social Science Citation Index follows any comprehensible rationale'.

4.1 Community detection algorithms

Community detection involves trying to find distinctive groups within a network. It is effectively the application of cluster analysis to graphs. Whereas a cluster, roughly speaking, is a group of data points that are closer or more similar to each other than they are to points in other clusters, a community is a group of nodes with

more (or more highly weighted) links to each other than to nodes in other communities.

The topic of community detection has been comprehensively reviewed by Fortunato (2010), who provides a detailed survey of the entire field, and by Malliaros and Vazirgiannis (2013), who pay special attention to community detection in directed networks. A much shorter, accessible summary is provided by Newman (2012).

Communities are also referred to as modules, groups or clusters. A community that induces a complete subgraph—where every member node shares a link with every other member node—is called a clique. A division of a graph into non-overlapping clusters, such that each vertex belongs to one cluster, is called a partition. The fuzzy analogue to a partition, where a vertex may belong to multiple overlapping communities, is called a cover. Sequential partitions form a hierarchical clustering structure that may be represented on a dendrogram.

A highly specialised case of community detection is *graph partitioning*, where the vertices of a graph are divided into g groups of predefined size (by some scheme, e.g. that every group is of equal size) such that the number of inter-group edges is minimised. Though this method sees applications in computing, circuit design and linear algebra (Fortunato, 2010), in most network analysis problems the number and relative size of clusters is not known in advance.

Indeed, community detection is an ill-defined problem: there is no universal definition of a community or cluster. Unlike in classification problems, a "ground truth" community structure typically does not exist. Therefore typically the main goal of community detection is simply to find a way of summarising network data in a convenient and useful form.

This chapter looks at a few popular methods for community detection in networks. It is not, nor could it ever be, an exhaustive review of the subject, which is vast.

To give a flavour of the different results yielded by different community detection methods, a selection of the algorithms have been applied to a simple dataset: the citation network of 47 statistical journals introduced by Varin et al. (2016)[2]. The results are discussed in Section 4.2.

[2] These data are readily accessible from R package `scrooge` by calling the command `data(citations)`. Reproducible code for these analyses are available in a supplementary vignette `http://selbydavid.com/vignettes/statsclustering.html`.

4.1.1 Traditional clustering algorithms

Cluster analysis involves dividing objects into groups so that objects in the same group are similar and those in different groups are different. Differences between data points are quantified by a proximity measure, such as the Minkowsi distance, cosine similarity or Pearson correlation coefficient. Traditional clustering algorithms are generally divided into partitional and hierarchical clustering.

Partitional algorithms, the most famous of which is k-means clustering, typically divide the data points randomly into k groups,

then perform expectation-maximisation iterations moving points between the groups to optimise an objective function. Points can be hard-assigned to groups or they can have fractional membership, which may be calculated from the likelihoods under a mixture model.

Hierarchical algorithms do not require the number of clusters to be specified in advance. We start with every point in its own group (or all points in one group). Then, using the chosen distance metric and a linkage criterion (a method of defining distances between groups of points), groups are successively agglomerated or divided until all points are in the same group or all are in different groups. The sequence of partitions describes a hierarchical structure, and can be represented on a dendrogram, like Figure 4.1.

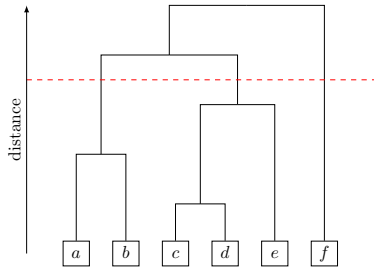


Figure 4.1: An example dendrogram. Cutting the tree along the dashed line will partition these data into three clusters: $\{a, b\}$, $\{c, d, e\}$ and $\{f\}$.

Before performing hierarchical or partitional clustering, we can also preprocess the data using spectral analysis. Such an approach is called spectral clustering. The original data matrix is transformed into a set of points in space represented by eigenvectors. These coordinates are then grouped using standard hierarchical or partitional clustering. For instance, if we want to find k clusters, we compute the k lowest eigenvalues of the Laplacian matrix. We then build an $n \times k$ matrix $\mathbf{V}$, comprising the k eigenvectors. This represents the data points in k-dimensional Euclidean space and Cartesian coordinates. We can then apply k-means or other algorithms. *Normalised* spectral clustering uses the normalised Lapacian and also normalises the matrix $\mathbf{V}$ by dividing each row by its sum (Fortunato, 2010).

When presented with network data, one approach to community detection is to consider the adjacency matrix like any n-dimensional data set, treating the nodes as n data points and choosing a traditional clustering algorithm to group them.

An advantage of this approach is the simplicity of implementation: a range of popular data clustering algorithms are available in standard statistical software packages. A recent example is Varin et al. (2016), who used hierarchical clustering and Pearson correlation distance of citation counts to cluster statistical journals—reproduced in Table 4.1.

When applying conventional data clustering algorithms to network data, however, there is no guarantee that the clusters returned will be internally connected. This can be a problem on sparse graphs; indeed, applying the same method as Varin et al. (2016) to a larger citation network results in some clusters containing disjoint components. One fix might be to measure distance by simply counting the number of paths between pairs of nodes, however this method has its own problems (Girvan and Newman, 2002).

In the next sections we will explore several classes of dedicated community detection algorithms. For further information on traditional data clustering, good overviews are given by Xu and Wunsch (2008) or Everitt et al. (2011).

Group	Members
1	American Statistician, International Statistical Review
2	Annals Of The Institute Of Statistical Mathematics, Australian & New Zealand Journal Of Statistics, Communications In Statistics: Theory And Methods, Journal Of Statistical Planning And Inference, Journal Of Time Series Analysis, Metrika, Statistics, Statistical Papers, Statistics & Probability Letters
3	Annals Of Statistics, Bernoulli, Biometrika, Canadian Journal Of Statistics: Revue Canadienne De Statistique, Journal Of The American Statistical Association, Journal Of Computational And Graphical Statistics, Journal Of Multivariate Analysis, Journal Of Nonparametric Statistics, Journal Of The Royal Statistical Society Series B: Statistical Methodology, Scandinavian Journal Of Statistics, Statistics And Computing, Statistica Neerlandica, Statistica Sinica, Test
4	Biometrical Journal, Biometrics, Biostatistics, Journal Of Biopharmaceutical Statistics, Journal Of The Royal Statistical Society Series A: Statistics In Society, Journal Of The Royal Statistical Society Series C: Applied Statistics, Lifetime Data Analysis, Statistics In Medicine, Statistical Methods In Medical Research, Statistical Modelling, Statistical Science
5	Communications In Statistics: Simulation And Computation, Computational Statistics, Computational Statistics & Data Analysis, Journal Of Applied Statistics, Journal Of Statistical Computation And Simulation, Technometrics
6	Environmental And Ecological Statistics, Environmetrics, Journal Of Agricultural Biological And Environmental Statistics
7	Journal Of Statistical Software
8	Stata Journal

Table 4.1: A grouping of 47 statistics journals, using the same agglomerative hierarchical clustering approach as Varin et al. (2016)

4.1.2 *Modularity optimisation*

Girvan and Newman (2002) proposed a divisive hierarchical clustering algorithm that involves successively removing edges with the highest *betweenness* in a graph. An edge's betweenness is the number of shortest paths between vertices that run along it. The algorithm works as follows.

1. Calculate the betweenness of every edge in the graph.
2. Remove the edge with the highest betweenness.
3. Recalculate betweenness for any edges where it may have changed.
4. Repeat steps 2 and 3 until there are no edges left.

The procedure seems intuitive: communities are groups of vertices with more edges within than between them. Removing the latter should thus break the graph up into components representing these groups. Unlike traditional hierarchical clustering, every cluster yielded by the algorithm will surely be (internally) connected.

As this is a hierarchical clustering algorithm, there is no stopping criterion: the entire sequence of partitions (the dendrogram) is returned, with no obvious indication of where to cut the tree to retrieve a 'good' partition. This becomes more of a problem for large graphs whose clustering dendrogram may not be easy to visualise.

A later refinement by Newman and Girvan (2004) suggested selecting the partition (from the tree) that maximises *modularity*: the proportion of edges that are within communities, minus the

expected proportion of within-community edges if the graph were random with no community structure. Modularity is defined as

$$Q = \sum_i (e_{ii} - a_i^2), \tag{4.1}$$

where e_{ij} is the fraction of edges in the network that connect communities i and j, with $a_i = \sum_j e_{ij}$. The higher the value of modularity, the more evidence there is of non-null community structure.

The edge-betweenness algorithm has been widely applied in biology, sociology and computer science. It is implemented in a number of software libraries: for example, in R, the `igraph` network analysis package (Csardi and Nepusz, 2006) includes a function called `cluster_edge_betweenness`.

Running the method of Newman and Girvan (2004) on our statistical journals network obtains the grouping given in Table 4.2.

Despite its popularity, the computational complexity of the edge-betweenness algorithm—around $O(m^2 n)$ time, where m and n are the numbers of edges and vertices, respectively—limits it to networks of only a few thousand vertices. The citation data we are interested in include over ten thousand journals; if disaggregated, these publications comprise many hundreds of thousands or even millions of articles.

As modularity is a measure of quality of community structure, Newman (2004b) proposed an alternative approach: rather than iteratively remove edges with high edge betweenness, simply optimise the quantity (4.1) directly. A brute-force approach is computationally intractable, so instead they use a "greedy" optimisation in the form of agglomerative hierarchical clustering, merging communities that provide the largest increase in modularity. The greedy modularity optimisation algorithm has computational cost $O((m+n)n)$, or $O(n^2)$ on sparse graphs.

Clauset et al. (2004) further refined Newman's algorithm, using max heap data structures and other efficiency improvements, reducing complexity to $O(md \log n)$, where d is the depth of the dendrogram. Using this faster greedy algorithm, the authors were able to detect community structure in graphs with hundreds of thousands of vertices. The method is implemented in R's `igraph` package as `cluster_fast_greedy`, producing the output in Table 4.3.

Later, a group of researchers from Belgium's Université catholique de Louvain (Blondel et al., 2008) criticised the method of Clauset et al. (2004) because it sometimes produces partitions with modularity much lower than the results from alternative approaches, such as simulated annealing. They also noted the fast and greedy algorithm's "tendency to produce super-communities that contain a large fraction of the nodes, even on synthetic networks that have no significant community structure".

The Louvain authors proposed their own greedy modularity optimisation algorithm, which runs in $O(m)$ time and can be applied to graphs with hundreds of millions of edges. It is sometimes referred to as the "Louvain method" (Leydesdorff et al., 2016) and is

Group	Members
1	American Statistician
2	Annals Of The Institute Of Statistical Mathematics
3	Annals Of Statistics, Biometrika, Computational Statistics & Data Analysis, Journal Of The American Statistical Association, Journal Of Multivariate Analysis, Journal Of The Royal Statistical Society Series B: Statistical Methodology, Journal Of Statistical Planning And Inference, Statistica Sinica
4	Australian & New Zealand Journal Of Statistics
5	Bernoulli
6	Biometrical Journal
7	Biometrics, Statistics In Medicine
8	Biostatistics
9	Canadian Journal Of Statistics: Revue Canadienne De Statistique
10	Communications In Statistics: Simulation And Computation
11	Communications In Statistics: Theory And Methods
12	Computational Statistics
13	Environmental And Ecological Statistics
14	Environmetrics
15	International Statistical Review
16	Journal Of Agricultural Biological And Environmental Statistics
17	Journal Of Applied Statistics
18	Journal Of Biopharmaceutical Statistics
19	Journal Of Computational And Graphical Statistics
20	Journal Of Nonparametric Statistics
21	Journal Of The Royal Statistical Society Series A: Statistics In Society
22	Journal Of The Royal Statistical Society Series C: Applied Statistics
23	Journal Of Statistical Computation And Simulation
24	Journal Of Statistical Software
25	Journal Of Time Series Analysis
26	Lifetime Data Analysis
27	Metrika
28	Scandinavian Journal Of Statistics
29	Stata Journal
30	Statistics And Computing
31	Statistics
32	Statistical Methods In Medical Research
33	Statistical Modelling
34	Statistica Neerlandica
35	Statistical Papers
36	Statistics & Probability Letters
37	Statistical Science
38	Technometrics
39	Test

Table 4.2: A grouping of 47 statistics journals, obtained by running the edge betweenness algorithm (Girvan and Newman, 2002) and selecting the partition that maximises modularity (Newman and Girvan, 2004)

Table 4.3: A grouping of 47 statistics journals yielded by greedy modularity maximisation (Clauset et al., 2004)

Group	Members
1	Annals Of The Institute Of Statistical Mathematics, Australian & New Zealand Journal Of Statistics, Communications In Statistics: Simulation And Computation, Communications In Statistics: Theory And Methods, Journal Of Applied Statistics, Journal Of Multivariate Analysis, Journal Of Nonparametric Statistics, Journal Of Statistical Computation And Simulation, Journal Of Statistical Planning And Inference, Journal Of Time Series Analysis, Metrika, Statistics, Statistica Neerlandica, Statistical Papers, Statistics & Probability Letters, Technometrics
2	Computational Statistics, Computational Statistics & Data Analysis, Journal Of Statistical Software
3	American Statistician, International Statistical Review
4	Environmental And Ecological Statistics, Environmetrics, Journal Of Agricultural Biological And Environmental Statistics
5	Annals Of Statistics, Bernoulli, Biometrika, Canadian Journal Of Statistics: Revue Canadienne De Statistique, Journal Of The American Statistical Association, Journal Of Computational And Graphical Statistics, Journal Of The Royal Statistical Society Series B: Statistical Methodology, Scandinavian Journal Of Statistics, Statistics And Computing, Statistica Sinica, Test
6	Biometrical Journal, Biometrics, Biostatistics, Journal Of Biopharmaceutical Statistics, Journal Of The Royal Statistical Society Series A: Statistics In Society, Journal Of The Royal Statistical Society Series C: Applied Statistics, Lifetime Data Analysis, Statistics In Medicine, Statistical Methods In Medical Research, Statistical Modelling, Statistical Science
7	Stata Journal

implemented in `igraph` as `cluster_louvain`.

In the Louvain method, every node starts in its own (singleton) community. The algorithm proceeds in two phases.

1. For each node i: for each neighbouring[3] node j, calculate the change in modularity yielded by removing i from its own community to the community of node j. Move i to the community that would give the largest positive increase in modularity. If no move would yield an increase, leave i where it is. Repeat until no further moves are possible.
2. Construct a new network whose nodes are the communities found in phase 1. That is, aggregate all nodes in each community into a single 'super-node' representing that community. Within-community links become node self-loops; between-community links are aggregated into (weighted) inter-super-node links.

[3] In network terminology, the *neighbourhood* of node i is the induced subgraph of nodes adjacent (connected by an edge) to i.

The resulting network from phase 2 is then fed back into phase 1 and the process iterates.

Table 4.4 shows an example output, from applying the Louvain method to the statistics journals dataset of Varin et al. (2016).

By the late 2000s, modularity optimisation became widely known and had become "by far the most popular" approach to community detection in networks (Good et al., 2010; Fortunato, 2010). However, some authors have expressed reservations about its use in practice.

Though the Louvain method yields better modularity maxima than the algorithm of Clauset et al. (2004), Fortunato (2010) argues that forming communities around neighbourhoods of nodes (the first phase) may lead to "spurious" results. Fortunato (2010) con-

Group	Members
1	Environmental And Ecological Statistics, Environmetrics, Journal Of Agricultural Biological And Environmental Statistics
2	Computational Statistics, Computational Statistics & Data Analysis, Journal Of Statistical Software
3	Annals Of The Institute Of Statistical Mathematics, Australian & New Zealand Journal Of Statistics, Communications In Statistics: Simulation And Computation, Communications In Statistics: Theory And Methods, Journal Of Applied Statistics, Journal Of Multivariate Analysis, Journal Of Nonparametric Statistics, Journal Of Statistical Computation And Simulation, Journal Of Statistical Planning And Inference, Journal Of Time Series Analysis, Metrika, Statistics, Statistica Neerlandica, Statistical Papers, Statistics & Probability Letters, Technometrics
4	Stata Journal
5	Biometrical Journal, Biometrics, Biostatistics, Journal Of Biopharmaceutical Statistics, Journal Of The Royal Statistical Society Series A: Statistics In Society, Journal Of The Royal Statistical Society Series C: Applied Statistics, Lifetime Data Analysis, Statistics In Medicine, Statistical Methods In Medical Research, Statistical Modelling, Statistical Science
6	American Statistician, International Statistical Review
7	Annals Of Statistics, Bernoulli, Biometrika, Canadian Journal Of Statistics: Revue Canadienne De Statistique, Journal Of The American Statistical Association, Journal Of Computational And Graphical Statistics, Journal Of The Royal Statistical Society Series B: Statistical Methodology, Scandinavian Journal Of Statistics, Statistics And Computing, Statistica Sinica, Test

Table 4.4: A grouping of 47 statistics journals yielded by the 'Louvain method' (Blondel et al., 2008) community detection algorithm

cludes: "the accuracy of greedy optimization is not that good, as compared with other techniques".

Fortunato and Barthelemy (2007) demonstrated that modularity optimisation has a "resolution limit", meaning it generally fails to identify clusters smaller than size $\sqrt{2m}$, where m is the total number of edges in the graph.

The modularity function (4.1) assumes a random graph model where the expected number of links between communities depends on m. It tends to favour communities with similar total degree (Rosvall and Bergstrom, 2007). Good et al. (2010) showed that if communities are joined to each other by a constant number of edges, they will become increasingly likely to be merged together as the total number of communities increases, even if their internal structure does not change. An example is given in Figure 4.2.

As a result, the partition of the graph with highest modularity may depend more on the size of the graph rather than on any actual structure within it.

The same authors point out a "degeneracy" problem: often there are a very large number of alternative partitions with modularity close to the optimum. Therefore, the output of modularity optimisation algorithms "should be treated with caution in all but the most straightforward cases" (Good et al., 2010).

Leskovec et al. (2009) raise the question of whether community structure is detectable at all—let alone by modularity optimisation—in very large graphs. Instead of modularity, one can compute *conductance*, the ratio of a community's between-group edges to its

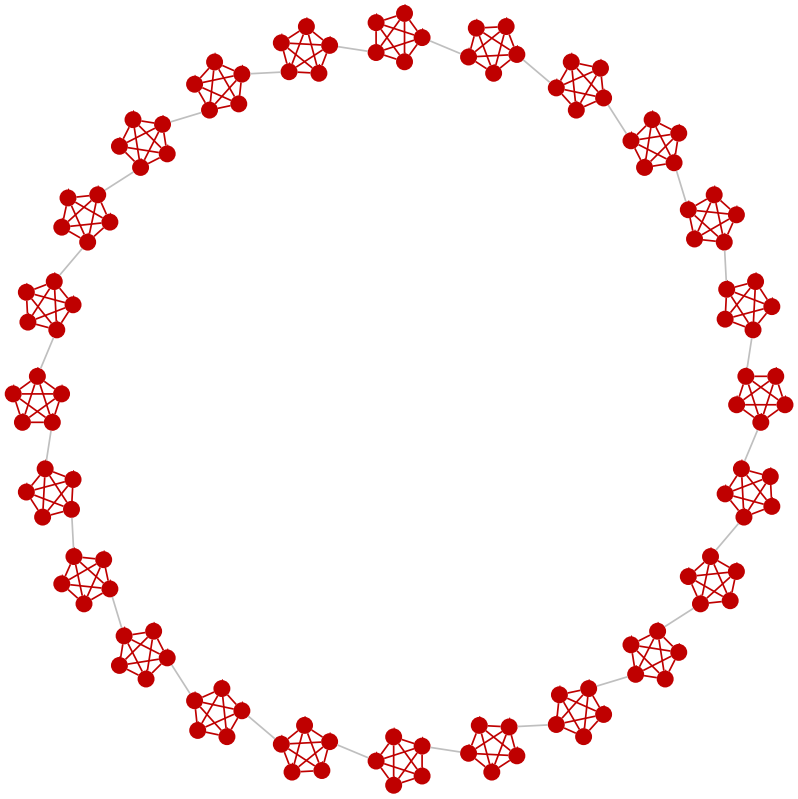


Figure 4.2: In this circular network there are 24 cliques comprising 5 nodes each, joined to each neighbouring clique by a single edge. Intuitively, we should have one clique per community, but the maximum modularity solution is to partition the graph into 12 pairs of adjacent cliques. Based on an example by Good et al. (2010)

within-group edges. The *network community profile plot*—the best possible value of conductance for a given community size—tends to rise for subgraphs larger than 100 vertices, implying that clusters are well-defined only when they are fairly small. However, this may be an artefact of conductance rather than a fundamental property of networks in general.

Visualisation of similarities (VOS) is an alternative to multidimensional scaling. According to van Eck and Waltman (2007), who first proposed the method, "the aim of VOS is to provide a low-dimensional visualization in which objects are located in such a way that the distance between any pair of objects reflects their similarity as accurately as possible". The method has been applied especially by bibliometricians to construct maps of citation networks (van Eck et al., 2010a; Waltman et al., 2010; Leydesdorff et al., 2016).

Suppose we have a similarity matrix, $\mathbf{S} = (s_{ij})_{n\times n}$, where $s_{ij} \geq 0$, $s_{ij} = s_{ji}$ and $s_{ii} = 0$ for all $i, j = 1, \ldots, n$. Let $\mathbf{X} = (x_{ij})_{n\times m}$ denote a matrix of coordinates (to be determined) in m-dimensional space, where x_{ij} is the j^{th} coordinate of object i. Then VOS minimises the objective function

$$V(\mathbf{X}; \mathbf{S}) = \sum_{i<j} s_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2, \tag{4.2}$$

subject to the constraint

$$\sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\| = \frac{n(n-1)}{2}, \tag{4.3}$$

where $\|\cdot\|$ denotes the Euclidean norm. The constraint (4.3) is imposed to avoid a trivial solution where every object is given the same coordinates, $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n$.

Under certain conditions, visualisation of similarities is equivalent to a variant of multidimensional scaling called *Sammon mapping* (van Eck and Waltman, 2007; van Eck et al., 2010a). In particular, minimising the objective function (4.2) subject to the constraint (4.3) yields the same solution as minimising the unconstrained function

$$V_u(\mathbf{X}; \mathbf{S}) = \sum_{i<j} s_{ij} d_{ij}^2 - \sum_{i<j} d_{ij}, \tag{4.4}$$

where $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$.

Unlike standard multidimensional scaling or principal coordinates analysis, visualisation of similarities does not exhibit visual artifacts like the "horseshoe effect" (van Eck and Waltman, 2007) or "circular maps" (van Eck et al., 2010b).

What relevance have these techniques to community detection?

Mapping and dimensionality reduction methods can be used in combination with clustering algorithms. Sometimes the approach is simply to use the map to visualise a partition that has been computed already. However, the map might also be used to cluster nodes based on their coordinates.

Waltman et al. (2010) proposed using (4.4) as part of a "unified approach to mapping and clustering of bibliometric networks". For mapping, d_{ij} is defined as above. When applied to clustering, we let $d_{ij} = 0$ if nodes i and j are in the same cluster and $d_{ij} = 1/\gamma$ otherwise. The parameter $\gamma > 0$ is called the *resolution parameter*; the larger the value of γ, the larger the number of clusters that may be obtained.

Interestingly, Waltman et al. (2010) showed that VOS clustering is a generalisation of modularity optimisation (Section 4.1.2). The modularity function of Newman and Girvan (2004) is equivalent to (4.4) when the resolution parameter is $\gamma = 1$ and when the similarity matrix is defined by association strength—also known as the proximity or probability affinity index (see van Eck and Waltman, 2009).

Increasing the value of the resolution parameter can help tackle the "resolution limit problem" (Fortunato and Barthelemy, 2007) of modularity-based community detection, which otherwise may fail to identify relatively small-sized clusters in the network (Waltman et al., 2010).

4.1.3 *Information-theoretic methods*

Modularity-based community detection algorithms, though popular, face a number of practical problems, as mentioned in Section 4.1.2.

Rosvall and Bergstrom (2007) proposed *cluster compression*, an information-theoretic method that considers the conditional in-

formation $H(X|Y)$, the information necessary to describe X, the original network, given Y, a simplified description of it.

The basic idea is to treat community detection as a communication process. Person A knows the detailed structure of the original graph and wants to transmit this information as a compressed message to person B. Ideally, we strike a balance between the minimising the length of the message and maximising the amount of information contained within it.

Simply minimising $H(X|Y)$ will lead to the trivial solution where $X = Y$, thus cluster compression actually seeks the *minimum description length* (Rissanen, 1978),

$$L(Y) + L(X|Y) = n \log k + \frac{1}{2} k(k+1) \log l + H(X|Y) \qquad (4.5)$$

where $L(Y)$ is the length in bits of the message, $L(X|Y)$ is the number of bits of additional information required to reconstruct X exactly, k is the number of clusters, l is the total number of edges and n is the total number of vertices in the graph.

It is not computationally feasible to evaluate this quantity for every possible partition of the network, so Rosvall and Bergstrom (2007) suggest using simulated annealing with the heat-bath algorithm to find the optimum. They were able to apply this method to networks as large as 10,000 nodes.

Rosvall and Bergstrom (2008) took the idea of code compression for community detection further, with another information-theoretic method known as the *Infomap* algorithm, based on the "map equation" framework. The method was first proposed by Rosvall and Bergstrom (2008); a later paper by Rosvall et al. (2009) covers the topic in greater detail.

The "map" terminology refers to street maps: street names are unique within but not between cities, while the city names are themselves are usually unique. Thus we can reuse street names between cities without causing confusion: if I live in Coventry and a neighbour arranges to meet me on the High Street, they are unlikely to be referring to the one in Edinburgh, for example.

We can model the community detection problem as a coding problem where we aim to describe—as succinctly as possible—the trajectory of a random walker around the graph. The shortest codewords are given to the most commonly visited nodes and longer codewords are given to rarer ones, a lossless compression scheme known as *Huffman coding*. Modules (cities) are then introduced as an extra codebook. Every module is uniquely coded, but only referred to when moving from one module to another (once the random walker is sent to Coventry, they are assumed to stay in that city until specified otherwise). If the communities are well separated then moves between them will be infrequent, thus by reusing node codewords between clusters, a two-level description will be considerably shorter than a one-level description.

The map equation is

$$L(\mathsf{M}) = q_{\curvearrowright} H(\mathcal{Q}) + \sum_{i=1}^{m} p^i_{\circlearrowright} H(\mathcal{P}^i), \tag{4.6}$$

where $L(\mathsf{M})$ is the average number of bits per step to describe a random walk on a network with partitioning M. Here, $H(\mathcal{Q})$ is the entropy of movements between modules, $H(\mathcal{P}^i)$ is the entropy of movements within module i, $q_{\curvearrowright}$ is the probability of moving between modules and $p^i_{\circlearrowright}$ is probability of moving within module i (plus the probability of leaving it).

The algorithm works as follows.

1. Compute the ergodic node visit frequencies—that is, the damped PageRank scores—of every node in the network. Assign every node to its own module (cluster). Compute the exit probability for each module, given by

$$q_{i\curvearrowright} = \tau \frac{n - n_i}{n - 1} \sum_{\alpha \in i} p_\alpha + (1 - \tau) \sum_{\alpha \in i} \sum_{\beta \notin i} p_\alpha w_{\alpha\beta}, \tag{4.7}$$

   where i is the index of a module, τ is a damping factor (representing a teleportation term), n is the total number of nodes, n_i is the number of nodes in module i, α is some node within module i, p_α is the PageRank score of node α, β is some node in another module and $w_{\alpha\beta}$ is the weight of outgoing links from α to β.

2. Perform a *greedy search*: use the map equation (4.6) to compute the average bits per step of the current partition. Then merge the two modules which give the largest decrease in description length. Repeat this process until the description length can no longer decrease.

3. Refine the result of the greedy search using simulated annealing and the heat-bath algorithm, initialised by the partition given by the greedy search. Run several times with different temperatures and choose the run that gives the shortest description (minimising the map equation).

Compared with modularity optimisation (Section 4.1.2), Rosvall and Bergstrom (2008; 2009) suggest that the map equation approach is better at detecting structure in networks where links represent "patterns of movement among nodes"—such as citation networks—whereas modularity-based methods may be preferred for networks where links represent pairwise relationships, such as social networks.

When applied to the statistics journals dataset of Varin et al. (2016), the Infomap algorithm yields the grouping given in Table 4.5.

4.1.4 Overlapping communities

So far, this chapter has only considered partitions—divisions of the data into non-overlapping communities or clusters. Real life

Group	Members
1	Annals Of The Institute Of Statistical Mathematics, Biometrika, Journal Of The American Statistical Association, Journal Of Biopharmaceutical Statistics, Journal Of Nonparametric Statistics, Journal Of The Royal Statistical Society Series B: Statistical Methodology, Journal Of Statistical Computation And Simulation, Metrika, Scandinavian Journal Of Statistics, Stata Journal, Technometrics
2	Annals Of Statistics, Australian & New Zealand Journal Of Statistics, Communications In Statistics: Theory And Methods, Environmental And Ecological Statistics, Journal Of Multivariate Analysis, Journal Of The Royal Statistical Society Series C: Applied Statistics, Journal Of Statistical Planning And Inference, Statistical Methods In Medical Research, Statistical Science, Test
3	Biometrics, Biostatistics, Communications In Statistics: Simulation And Computation, International Statistical Review, Journal Of Computational And Graphical Statistics, Journal Of The Royal Statistical Society Series A: Statistics In Society, Journal Of Time Series Analysis, Lifetime Data Analysis, Statistics And Computing, Statistical Papers, Statistics & Probability Letters
4	Bernoulli, Biometrical Journal, Computational Statistics, Journal Of Statistical Software, Statistical Modelling, Statistica Sinica
5	American Statistician, Canadian Journal Of Statistics: Revue Canadienne De Statistique, Computational Statistics & Data Analysis, Journal Of Applied Statistics, Statistics
6	Environmetrics, Journal Of Agricultural Biological And Environmental Statistics, Statistics In Medicine, Statistica Neerlandica

Table 4.5: A grouping of 47 statistics journals obtained using one run of the Infomap algorithm (Rosvall and Bergstrom, 2008). The method is partly nondeterministic, so it may not always return this exact output

is rarely so clear-cut: if our main task is categorising academic publications by field, then multidisciplinary journals will throw a spanner in the works. A journal like *Biostatistics* may belong to biology and to statistics; a non-overlapping cluster structure may put it in one field but not the other, or it may merge the two fields.

In the Bayesian nonparametrics literature, this concept is known as *feature allocation*, where discrete clusters are generalized as non-integer 'features' or 'topics', and each data point may belong to an arbitrary number of them (Broderick et al., 2013).

According to Fortunato (2010), the most popular overlapping community detection technique is the *clique percolation method* (Palla et al., 2005), an approach applicable to graphs with up to 10^5 vertices.

A *clique* is group of nodes that induces a complete subgraph; that is, where each node is connected to every other node in the group. Clique percolation works on the basis that communities usually comprise several such groups connected together by sharing nodes. A k-clique is a clique with k nodes in it (a complete subgraph of size k). Two k-cliques are adjacent if they share $k - 1$ nodes. A k-clique community is defined to be the union of all k-cliques reachable from each other via adjacent k-cliques.

Palla et al. (2005) propose using an exponential-time algorithm to retrieve the cover of k-clique communities in a graph. The procedure involves first locating all maximal cliques, and constructing a "clique–clique overlap matrix": a symmetric contingency table of the number of common nodes in each pair of cliques. The diagonal entries of the matrix are equal to the number of nodes in each clique. For a chosen value of k, all off-diagonal elements smaller

than $k-1$ and all diagonal elements smaller than k are set equal to zero. The remaining values are set equal to 1. Then the k-clique communities are the connected components of the resulting binary adjacency matrix.

Fortunato (2010) notes: "an interesting aspect of k-clique communities is that they allow to make a clear distinction between random graphs and graphs with community structure", unlike modularity (Section 4.1.2) which can take large values even when community structure does not exist.

On the other hand, k-clique communities may not be the best way to summarise a graph's structure, because any nodes not in a k-clique will not be in a k-clique community. For $k > 2$, all leaves (nodes connected to the rest of a graph by one edge) will be ignored, which could represent a large proportion of the network. Thus clique percolation does not actually return a *cover* of the graph, where every vertex is assigned to at least one cluster.

In its original form, clique percolation is defined for undirected, unweighted graphs. Although extensions to weighted and directed networks have been proposed, these tend to involve somewhat arbitrary workarounds, such as replacing all weights greater than a certain value with ones, setting all other weights to zero and then treating the graph as unweighted (Fortunato, 2010). Nonetheless, Palla et al. (2005) provide guidelines on choosing sensible values of k and the level at which to threshold the weights.

Cliques can be retrieved in `igraph` using the `cliques` function. An R implementation of the clique percolation method has been written by Angelo Salatino and is available on GitHub[4].

[4] `https://github.com/angelosalatino/CliquePercolationMethod-R`

Though the clique percolation method just described can find overlapping communities, it does not reveal hierarchical structure. Lancichinetti et al. (2009) proposed the first algorithm to detect communities that are both overlapping and hierarchical. It has worst-case computational complexity of $O(n^2 \log n)$. The basic idea is to optimise an objective function that will yield overlapping communities, then adjust a resolution parameter (c.f. VOS) to construct a hierarchy of the communities at different levels of resolution.

Quality or "fitness" of community i is measured by the function

$$Q(i) = \frac{2e_{ii}}{\left(2e_{ii} + \sum_j e_{ij}\right)^{\alpha}}, \tag{4.8}$$

where e_{ii} is the number of edges between nodes in community i and $\sum_j e_{ij}$ is the total number of edges connecting community i with the rest of the graph. The resolution parameter, α, is a positive real number that controls the size of communities.

The global maximum of this function is simply the entire graph, because such an all-encompassing community would have no external edges. Instead of this trivial solution, we seek *local* optima describing the so-called *natural community* for each node.

Given a node x, its natural community, $\mathcal{G}_x$, is calculated as follows.

1. Initially, let $\mathcal{G}_x = \{x\}$.
2. For every neighbouring node $y \notin \mathcal{G}_x$, calculate $Q(\mathcal{G}_x \cup y) - Q(\mathcal{G}_x)$, the change in fitness that would result from adding the neighbour to the community.
3. Add the neighbour that yields the largest positive increase in fitness. If no node yields an increase then stop.
4. For each node y now in $\mathcal{G}_x$, calculate $Q(\mathcal{G}_x \setminus y) - Q(\mathcal{G}_x)$ the change in fitness that would result in removing that node from the community.
5. If, for a particular node in $\mathcal{G}_x$ this change is positive, remove that node, then go to step 4. Otherwise go to step 2.

One way to compute a complete cover of the network would be to run the above algorithm for each and every node, but this would be computationally intensive. A more efficient (possibly less accurate) approach is to do the following.

1. Pick a random node, say x, that has not yet been assigned to any group.
2. Use the algorithm above to detect the natural community of x. Any node can be added to the group $\mathcal{G}_x$, regardless of whether it is already in another group or not.
3. Repeat from 1.

According to Lancichinetti et al. (2009), the loss in accuracy from not initialising at every single node is "minimal".

These procedures give us overlapping communities—several different sets of which, if computed with multiple different values of the resolution parameter α. It remains to explore their hierarchical structure.

A partition $\mathcal{C}'$ is said to be *hierarchically ordered* above partition $\mathcal{C}''$ if there exists a single community in $\mathcal{C}'$ that contains every member of $\mathcal{C}''$. (Communities can overlap, so this community of $\mathcal{C}'$ need not be the *only* community to which the members of $\mathcal{C}''$ belong.)

Empirical analysis by Lancichinetti et al. (2009) found that their algorithm outperforms the clique percolation method on many classical community detection benchmarks, but that clique percolation works better on networks containing many cliques.

Unlike clique percolation, the overlapping-hierarchical framework may be extended to weighted networks without arbitrary thresholds: simply replace the edge counts in (4.8) with the sums of the respective edge weights. An extension to directed networks has also been mooted, but not tested; this would involve considering the in-degree of nodes from outside the community rather than simply the the number of edges linking internal and external nodes.

Lancichinetti et al. emphasise that the abovementioned algorithm is a specific case of a general framework for recovering hierarchical communities; the fitness equation (4.8) can be replaced with any other objective function with a tunable resolution parameter, such as (4.4). The wider class of such methods is called *multiresolution methods* (Fortunato, 2010, p63).

Gregory (2009) introduced a different approach that aimed to reduce the dichotomy between "disjoint" and "overlapping" community detection algorithms. Rather than use a dedicated algorithm to detect overlapping communities (as in the previous two sections), Gregory proposed exploiting conventional "disjoint" community detection algorithms in a special way so that they can yield overlapping communities.

The method, called "Peacock", works by transforming the network into a larger one to which the conventional community detection algorithm can be applied. The partition yielded by this procedure is then transformed into (possibly overlapping) communities corresponding to the original network.

Peacock is inspired by another algorithm called "Conga" or "Congo" (Gregory, 2008), itself based on Girvan and Newman's (2002) edge betweenness algorithm. Each vertex may be split, based on a quantity called *split betweenness*, into two vertices and an edge between them. The Peacock algorithm successively splits the vertices with highest split betweenness, while remembering the original vertices from which they were generated. The transformed network is then fed into any conventional community detection algorithm. Suppose a vertex v is split into two vertices, v' and v''. Then, if these latter vertices are assigned to two different communities, this means the original vertex v belongs to both those communities.

Though Peacock is not implemented in `igraph`, it should be relatively straightforward to exploit the `betweenness` or `estimate_betweenness` functions to do so.

4.1.5 *Stochastic block modelling*[5]

[5] Though it is common in the literature, I am not especially convinced it is necessary to make "blockmodel" one word, versus "block model" or "block-model".

None of the community detection methods mentioned so far is particularly 'statistical'—even the null random graph model from the modularity function (4.1) is something of a straw man. Indeed, many numerical approaches to network analysis lack any quantification of uncertainty, providing descriptive statistics only (Snijders and Borgatti, 1999). But community detection can be considered as a statistical inference problem.

Block modelling is a method—originally deterministic—of permuting rows and columns in a network's adjacency matrix to reveal patterns in the structure. Where the block pattern is only approximate, a *statistical* block model allows quantification of its goodness of fit. Anderson et al. (1992) define a *stochastic block model* to be a probability distribution over graphs, where vertices are grouped into blocks. The probability of an edge existing between two vertices depends entirely on the blocks to which they belong. If blocks are unknown—as in community detection—then the method is called *a posteriori block modelling* (Snijders and Nowicki, 1997). Block modelling may be applied to directed as well as undirected graphs.

The stochastic block model of Anderson et al. (1992) is based on

the so-called p_1 distribution (Holland and Leinhardt, 1981). Given X, the adjacency matrix of a (binary) network,

$$p_1(x) = P(X = x) \propto \exp\left\{\rho m + \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j}\right\}, \quad (4.9)$$

where m is the number of mutual/reciprocated links, $x_{i+} = \sum_j x_{ij}$ is the out-degree of node i, x_{+j} is the in-degree of node j and ρ, θ, α_i and β_j are parameters corresponding to reciprocity, density, productivity and attractiveness, respectively. Given a block partition, the p_1 blockmodel involves aggregating the nodes in each block and then fitting the model (via least squares) to the resulting aggregated network.

Anderson et al. (1992) suggested fitting the p_1 model to the entire network, then comparing the parameters visually (or using cluster analysis; see Section 4.1.1) to group nodes into blocks. As Snijders and Nowicki (1997) emphasise, such methods are unreliable when the p_1 model does not fit the disaggregated data well.

Instead, Snijders and Nowicki (1997) recommended a "coloured graph model", or random block model. Suppose nodes can have one of m different colours, which correspond to blocks; these are treated as random variables. The probabilities of each colour are given by the vector $\theta = (\theta_1, \ldots, \theta_m)$ and the colour-conditional Bernoulli probabilities of edges between blocks are given by the matrix $\eta = (\eta_{kl})_{m \times m}$. These $m(m+3)/2$ parameters may be estimated using Gibbs sampling.

Newman (2012) points out,

> "... perhaps the most promising feature of the blockmodel method is that it is not limited to detecting traditional community structure in networks. In principle, any type of structure that can be formulated as a probabilistic model can be detected, including overlapping communities, bipartite or k-partite structures, communities within communities and many others."

According to Karrer and Newman (2011) and Newman (2012), the performance of stochastic block models suffers when applied to networks with widely-varying degrees. A degree-corrected generalisation of stochastic block models was proposed by Karrer and Newman at the time; more recently Peng and Carvalho (2016) suggested a Bayesian degree-corrected approach. Another Bayesian MCMC implementation was given by McDaid et al. (2013).

4.1.6 Spin glass model

Community detection can be formulated as a spin glass problem in statistical mechanics. A spin glass[6] is a magnet comprising particles whose spins are not all aligned in the same direction. Though orientations of spins might appear random, there exists a particular combination of spins that yields the minimum of potential energy, called the ground state, into which the spins will settle. Neighbouring particles magnetically interact with one another by way

[6] The phrase "spin *glass*" is an analogy to glass, which has no crystalline structure and whose atoms seem to be positioned randomly. Stein and Newman (2012) provide an accessible introduction to spin glasses and why they are interesting to physicists and mathematicians.

of spin–spin coupling, also called dipolar interaction, with energy dependent on the coupling strength and spin states of each pair of points (Edwards and Anderson, 1975).

Reichardt and Bornholdt (2006) drew an analogy between particles in a spin glass and nodes in a network: the spin states correspond to community labels and the magnetic interactions represent the presence or absence of links between nodes. Maximising a quality measure, such as modularity, then becomes equivalent to minimising the energy of the spin glass system. Assuming undirected edges, the Hamiltonian (total energy of the system) of a Potts model is given by

$$\mathcal{H} = -\sum_{i \neq j} (A_{ij} - \gamma p_{ij}) \delta(\sigma_i, \sigma_j), \tag{4.10}$$

where A is the (weighted, due to Heimo et al., 2008) adjacency matrix, $\gamma > 0$ represents the influence of present/absent links, p_{ij} is the expected weight of links between i and j under a null model, σ_i is the spin state at node i and δ is the Kronecker delta function.

In this system, dipolar interaction is assumed to work over an infinite range. Where links in the network are present, the effect is ferromagnetic: that is, where spins align in the same direction. Where links in the network are absent, the effect is antiferromagnetic—the spins will tend to point in opposite directions.

Equation (4.10) is a generalisation of the modularity function in equation (4.1), thus modularity maximisation is equivalent to minimising the Hamiltonian (for a particular choice of γ and p_{ij}). Indeed, the parameter γ acts as a resolution parameter, controlling the size of clusters and overcoming modularity's resolution limit.

The spin glass method is implemented in `igraph` as `cluster_spinglass`, and has been applied in Table 4.6.

Whereas Reichardt and Bornholdt recommended simulated annealing to minimise the Hamiltonian, Hastings (2006), who also considered community detection as a Potts model, suggests using belief propagation to find the ground state. Hofman and Wiggins (2008) extended this approach by considering the parameters of equation (4.10), such as the coupling strengths, as random variables, making a constrainted stochastic block model (see previous section). With a weak prior on K, the number of clusters, the system becomes a probabilistic model selection problem, solved via variational Bayes.

4.1.7 Dynamic communities

Community detection on time-dependent graphs is "still in its infancy" according to Fortunato (2010), and there is a "dearth of timestamped data on real graphs". Moreover, static community detection is far from solved—as can seen by the sheer variety of methods in the previous sections—so it might be premature to add another layer of complexity.

Group	Members
1	American Statistician, Biometrical Journal, Environmental And Ecological Statistics, International Statistical Review, Journal Of Applied Statistics, Journal Of Biopharmaceutical Statistics, Journal Of The Royal Statistical Society Series A: Statistics In Society, Journal Of The Royal Statistical Society Series C: Applied Statistics, Journal Of Statistical Software, Lifetime Data Analysis, Scandinavian Journal Of Statistics, Stata Journal, Statistics In Medicine, Statistical Methods In Medical Research, Statistical Modelling, Statistical Science
2	Annals Of The Institute Of Statistical Mathematics, Communications In Statistics: Simulation And Computation, Communications In Statistics: Theory And Methods, Computational Statistics, Computational Statistics & Data Analysis, Journal Of Multivariate Analysis, Journal Of Nonparametric Statistics, Journal Of Statistical Computation And Simulation, Journal Of Statistical Planning And Inference, Journal Of Time Series Analysis, Metrika, Statistics, Statistical Papers, Statistics & Probability Letters, Statistica Sinica, Technometrics, Test
3	Annals Of Statistics, Australian & New Zealand Journal Of Statistics, Bernoulli, Biometrics, Biometrika, Biostatistics, Canadian Journal Of Statistics: Revue Canadienne De Statistique, Environmetrics, Journal Of Agricultural Biological And Environmental Statistics, Journal Of The American Statistical Association, Journal Of Computational And Graphical Statistics, Journal Of The Royal Statistical Society Series B: Statistical Methodology, Statistics And Computing, Statistica Neerlandica

Table 4.6: A grouping of 47 statistics journals obtained using the spinglass algorithm of Reichardt and Bornholdt (2006)

For our purposes, we have a fairly limited range of time-series data so will probably not concern ourselves with trying to investigate communities' change over time. It would be fascinating, nonetheless, to see if we could observe the emergence, growth, shrinkage, splitting and merging of academic fields (or at least their journal-community proxies) over the years. This is the aim of dynamic topic modelling (Blei and Lafferty, 2006) based on the analysis of large text corpora. Examples of such models include Ahmed and Xing (2010) and Dubey et al. (2013). If looking at citation data only, an interesting example would be Stefaner (2009), who used the Infomap algorithm to plot the "changing nature of neuroscience".

4.1.8 Directed, weighted networks

Many of the community detection algorithms described above assume graphs with undirected, unweighted edges. But academic and social network graphs may be directed, weighted or both.

For example, friendships on Facebook are both undirected and unweighted: if I am your friend, then you are my friend and we are either friends or not. Relationships on Twitter are directed and unweighted: if I follow you, you need not reciprocate, and everyone who does follow you does so equally[7]. Co-citation and co-authorship networks are undirected and weighted: if I co-authored five papers with you, then you co-authored five papers with me. Journal citation networks are directed and weighted: journal i might cite journal j once and journal k several times and neither need return the favour.

[7] Since around 2010, Facebook has also offered a 'follow' feature, letting users subscribe to other users' public updates without being friends.

In their review of community detection method for directed

networks, Malliaros and Vazirgiannis (2013) observed: "The most common way to dealing with edge directionality during the clustering task, is simply to ignore it". This is an oversimplification that can conceal structure in the graph, as demonstrated in Figure 4.3.

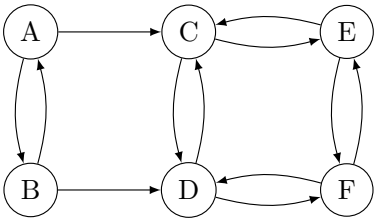


Figure 4.3: This graph can be divided into two communities: $\{A, B\}$ and $\{C, D, E, F\}$, as the former set of vertices is not reachable from the latter. Ignoring directionality means discarding this information, resulting in a graph with no visible community structure. Figure adapted from Malliaros and Vazirgiannis (2013)

Some community detection algorithms have been extended to deal with directed networks. Leicht and Newman (2008), for example, proposed a directed version of modularity. Methods such as Infomap (Rosvall and Bergstrom, 2008) natively work on graphs with directed edges. Other techniques require special transformations of the graph.

> Developing methods of community detection for directed graphs is a hard task. For instance, a directed graph is characterized by asymmetrical matrices (adjacency matrix, Laplacian, etc.), so spectral analysis is much more complex. Only a few techniques can be easily extended from the undirected to the directed case. Otherwise, the problem must be formulated from scratch.
>
> — Fortunato (2010)

Newman (2004a) explain how some algorithms, originally formulated for unweighted networks, may be extended to handle weighted edges.

4.2 *Empirical analysis*

The previous section reviewed a range of community detection methods, with a selection of them applied to the dataset of statistics journals from Varin et al. (2016). Tables 4.1–4.6 show the groupings of journals suggested by each of these algorithms.

Edge betweenness is not a practical approach; an Intel Core i7 desktop computer took nearly a minute to run the algorithm on the relatively small 47-journal citation network. (Our full Web of Science data set—not analysed here—comprises over 10,000 journals and 20 million citations.) The poor scaleability of edge betweenness is one of the motivations for alternative approaches such as greedy modularity optimisation, as described in Section 4.1.2.

So it is slow—but what of the results? Hardly useful: we obtain 39 mostly singleton groups, as shown in Table 4.2. This implementation of edge betweeness, like the original definition in Section 4.1.2, ignores edge direction, which could mean discarding a lot of otherwise-important information about citation behaviour.

Greedy modularity optimisation and the Louvain method are much quicker (taking only a second or so to run) and yield more helpful (and identical) output, which is quite similar to the results of agglomerative hierarchical clustering. Environmental statistics journals get their own group, as do those about biomedical statistics. Theoretical statistics journals, such as *Biometrika* and *JRSS-B*, are put together, while computational statistics journals form a cluster and *Stata Journal* is always isolated.

Infomap's results are harder to reconcile. *Stata Journal* and the *Journal of Biopharmaceutical Statistics* are mixed in with theoretical statistics journals, whilst the three environmental journals are not all kept together, and some of them are put with *Statistics in Medicine*. This illogical grouping may simply correspond to a local optimum in the map equation. Results from the spin glass algorithm are also difficult to reason with, and perhaps have aggregated the journals into too few groups to be easily interpretable.

Which algorithm gives the best results? This dataset is small enough to evaluate groupings through expert judgment alone. An academic statistician familiar with the literature—or at least the meanings of the journal titles[8]—can deduce that the the results of hierarchical clustering, greedy modularity optimisation and the Louvain method are fairly reasonable.

[8] *Biometrika* is an exception to this rule, as its title implies a connection with biometrics, but its publishers say it is 'primarily a theoretical statistics journal'.

However, we may not always have the benefit of domain expertise or prior knowledge about communities in networks, and expert judgment would struggle with larger networks of more than 100 or more nodes. Moreover, an *ad hoc* approach is not reproducible, even within the same analysis, if two 'experts' disagree on their choices of best grouping.

A naïve quantitative approach might be to choose the clustering that optimises a particular quality score, such as modularity—scores are given in Table 4.7. Unsurprisingly, the algorithms that directly maximise modularity—greedy optimisation and the Louvain method—yield the highest score. But modularity has its own problems, as described in Section 4.1.2.

Algorithm	Modularity
Greedy optimisation	27.7
Louvain method	27.7
Hierarchical clustering	27.0
Spin glass	21.8
Edge betweenness	16.4
Infomap	14.1

Table 4.7: Modularity scores (%) from community detection algorithms applied to citation data for 47 statistical journals

Even if it is assumed such a quality score gives a defensible relative ranking of the results, it does not reveal anything about any further structure in the data, without diving into *ad hoc* expert analyses once again.

The next section proposes a statistical framework for evaluating communities extracted from citation networks.

4.3 *Diagnostics for community detection*

Several different criteria are routinely used to measure success or stopping times for community detection algorithms, including modularity (Newman, 2004a) and derived measures, the map equation (Rosvall and Bergstrom, 2008) and total energy (Reichardt and Bornholdt, 2006). More recently-proposed metrics include Weighted

Community Clustering (Prat-Pérez et al., 2014) and those that consider quality as a function of community size (Leskovec et al., 2010).

These quantities are not necessarily valid for comparing the output from different algorithms. Fortunato (2007) claim that as long as the number of communities is not fixed in advance, 'using the optimization of quality functions to identify communities will be unjustified'.

More recently, Creusefond et al. (2016) identified 'contexts' or groups of graphs with common characteristics for which certain types of quality functions may provide better results, measured against known community structure. Biswas and Biswas (2016) proposed a framework called 'relative inclination towards accuracy', using multiple-criteria decision analysis to combine measurements of topological community quality (e.g. modularity) with measurements of accuracy relative to a ground-truth structure. Both of these approaches depend, however, on some set of known community labels, which in the case of fairly abstract definitions like academic fields, not be available.

In this section, we propose a statistical framework using an implied log-linear model to assess the quality of community detection results. Detecting lack of fit, outliers and unexplained structure is routinely done in generalised linear models by way of residual diagnostics, and here made possible for community detection using this framework. The technique can be applied to output that assigns nodes into groups, and allows analysis of particular nodes both quantitatively and visually.

4.3.1 Community profiles

Let's revisit the idea of community detection. Are we interested in clusters for their own sake? Are they even real? What are we actually measuring here? For what purpose?

We model a journal (or author, or other entity)'s *citation profile*, which is a stochastic (i.e. non-negative, unit-sum) vector representing the distribution of that journal's outgoing citations. It is the transition probability vector that describes one step of a random walk around the citation graph. Suppose we have a citation matrix $\mathbf{X}$, where x_{ij} is the number of citations from journal j to journal i. Then the citation profile of journal j is the j^{th} column of the transition matrix $\mathbf{P}$, where

$$p_{ij} = \frac{x_{ij}}{\sum_k x_{kj}}.$$

For example, suppose we have three journals, A, B and C. If half of the references in journal C's bibliography were to journal A, a quarter to B and the remainder to itself, then journal C's citation profile would be given by the vector

$$\mathbf{p}_C = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Given some grouping of journals into communities (not necessarily binary or non-overlapping) we then construct a *community*

profile. The most straightforward way of calculating a community profile is to find the vector (weighted) sum of citations issued by each community member, then scale this vector to sum to one. The weighting can be by community membership or could be deliberately biased, for example by article count or by PageRank (or some other measure of influence) to ensure that bigger or more influential journals have more or less effect on the nature of their communities' profiles.

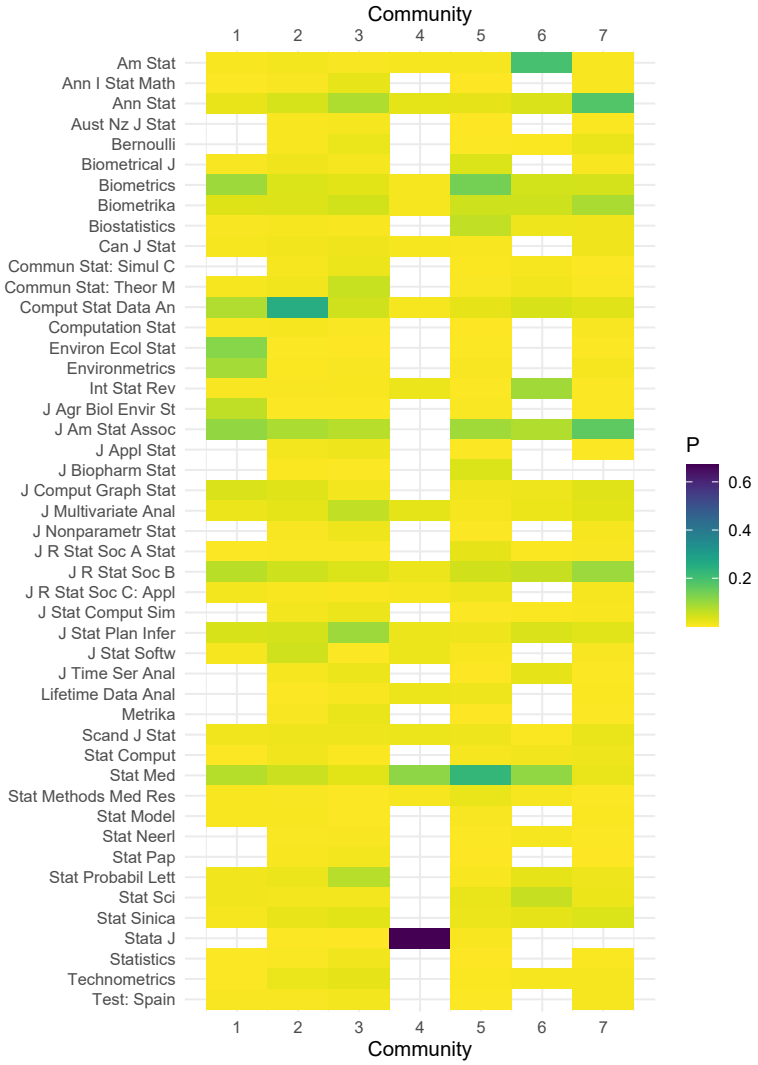


Figure 4.4: Community profile matrix heatmap for a clustering of journals via the Louvain method (given in Table 4.4)

Figure 4.4 represents a community profile matrix for the community structure obtained using the Louvain method of modularity optimisation. (A key to journal name abbreviations is given in Table 4.8.) From the graph it is readily apparent that over half of the citations from community 4 are to *Stata Journal*—i.e. itself—whereas communities 1 (environmental statistics) and 7 (statistical theory) do not cite *Stata Journal* at all. It is also easy to see that *Computational Statistics & Data Analysis* dominates the collective bibliography of community 2—of which it is a member.

Varin et al.	Journal Citation Reports	Full journal title
AmS	Am Stat	American Statistician
AISM	Ann I Stat Math	Annals Of The Institute Of Statistical Mathematics
AoS	Ann Stat	Annals Of Statistics
ANZS	Aust Nz J Stat	Australian & New Zealand Journal Of Statistics
Bern	Bernoulli	Bernoulli
BioJ	Biometrical J	Biometrical Journal
Bcs	Biometrics	Biometrics
Bka	Biometrika	Biometrika
Biost	Biostatistics	Biostatistics
CJS	Can J Stat	Canadian Journal Of Statistics: Revue Canadienne De Statistique
CSSC	Commun Stat: Simul C	Communications In Statistics: Simulation And Computation
CSTM	Commun Stat: Theor M	Communications In Statistics: Theory And Methods
CmpSt	Computation Stat	Computational Statistics
CSDA	Comput Stat Data An	Computational Statistics & Data Analysis
EES	Environ Ecol Stat	Environmental And Ecological Statistics
Envr	Environmetrics	Environmetrics
ISR	Int Stat Rev	International Statistical Review
JABES	J Agr Biol Envir St	Journal Of Agricultural Biological And Environmental Statistics
JASA	J Am Stat Assoc	Journal Of The American Statistical Association
JAS	J Appl Stat	Journal Of Applied Statistics
JBS	J Biopharm Stat	Journal Of Biopharmaceutical Statistics
JCGS	J Comput Graph Stat	Journal Of Computational And Graphical Statistics
JMA	J Multivariate Anal	Journal Of Multivariate Analysis
JNS	J Nonparametr Stat	Journal Of Nonparametric Statistics
JRSS-A	J R Stat Soc A Stat	Journal Of The Royal Statistical Society Series A: Statistics In Society
JRSS-B	J R Stat Soc B	Journal Of The Royal Statistical Society Series B: Statistical Methodology
JRSS-C	J R Stat Soc C: Appl	Journal Of The Royal Statistical Society Series C: Applied Statistics
JSCS	J Stat Comput Sim	Journal Of Statistical Computation And Simulation
JSPI	J Stat Plan Infer	Journal Of Statistical Planning And Inference
JSS	J Stat Softw	Journal Of Statistical Software
JTSA	J Time Ser Anal	Journal Of Time Series Analysis
LDA	Lifetime Data Anal	Lifetime Data Analysis
Mtka	Metrika	Metrika
SJS	Scand J Stat	Scandinavian Journal Of Statistics
StataJ	Stata J	Stata Journal
StCmp	Stat Comput	Statistics And Computing
Stats	Statistics	Statistics
StMed	Stat Med	Statistics In Medicine
SMMR	Stat Methods Med Res	Statistical Methods In Medical Research
StMod	Stat Model	Statistical Modelling
StNee	Stat Neerl	Statistica Neerlandica
StPap	Stat Pap	Statistical Papers
SPL	Stat Probabil Lett	Statistics & Probability Letters
StSci	Stat Sci	Statistical Science
StSin	Stat Sinica	Statistica Sinica
Tech	Technometrics	Technometrics
Test	Test: Spain	Test

Table 4.8: A key to different abbreviations of statistics journal titles. Full titles are according to Clarivate Analytics' Journal Citation Reports

4.3.2 *Convex hulls*

Assume community profiles are fixed and given. Then, which combination of communities best describes a particular journal's citation behaviour? The motivation behind this is that a field is determined by journals that are similar to each other, which—at least in the context of academic journals—can be observed through their outgoing citation behaviour. Though publications do *receive* as well as *issue* citations, authors and editors should (ignoring citation cartels) only have any control over the latter.

We seek the convex combinations of community profiles that are 'closest' (using some sensible definition of distance) to each individual journal's profile.

More concretely, let $\mathbb{S}^n$ denote the unit n-simplex: the space of non-negative real vectors of length $n+1$ whose elements sum to one. In symbolic terms, if $\mathbf{x} \in \mathbb{S}^{n-1}$ then

$$\mathbf{x} = \left\{ x_1, \ldots, x_n \;\middle|\; x_i \geq 0 \text{ and } \sum_{i=1}^{n} x_i = 1 \right\}.$$

For a citation network of n journals and k communities, let $\mathbf{j} \in \mathbb{S}^{n-1}$ denote a journal profile, let $\mathbf{c} \in \mathbb{S}^{n-1}$ denote a community profile, let $\lambda_1\mathbf{c}_1 + \cdots + \lambda_k\mathbf{c}_k$ denote a convex combination of community profiles (with weights $0 \leq \lambda_i \leq 1$) and let d denote a distance metric between two vectors, for example the Euclidean distance

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}.$$

Then the scalar quantity

$$d(\mathbf{j}, \lambda_1\mathbf{c}_1 + \cdots + \lambda_k\mathbf{c}_k)$$

is the distance between a journal profile and the convex hull of community profiles.

Statistically, we are interested in which journals/nodes are well described by the community profiles and which are not. Some aggregation of all the distances (weighted by citation counts) can be used as a measure of overall cluster quality.

By measuring the minimum distance from the convex hull of community profiles to an individual journal's citation profile (Figure 4.5), we can assess how well the community structure explains that journal's citation behaviour.

In a citation network of n journals, each community has a profile that lies in $\mathbf{S}^{n-1}$, because citation profiles—and their convex combinations—are stochastic n-vectors.

In vector notation, the (Euclidean) distance to minimise is

$$Q = d_2(\mathbf{j}, \mathbf{C}\lambda) = \mathbf{j}^T\mathbf{j} + \boldsymbol{\lambda}^T\mathbf{C}^T\mathbf{C}\boldsymbol{\lambda} - 2\mathbf{j}^T\mathbf{C}\boldsymbol{\lambda},$$

subject to

$$\mathbf{1}^T\boldsymbol{\lambda} = 1$$
$$\boldsymbol{\lambda} \geq \mathbf{0},$$

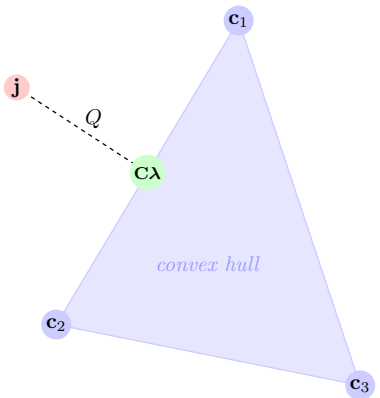


Figure 4.5: Our aim is to find $\boldsymbol{\lambda}$ so that $\mathbf{C}\boldsymbol{\lambda}$ is the 'closest' point on the convex hull of community profiles to $\mathbf{j}$, a given journal profile

which is equivalent to the quadratic programming problem of minimizing, with the same constraints,

$$Q_2 = \frac{1}{2}\boldsymbol{\lambda}^T\mathbf{C}^T\mathbf{C}\boldsymbol{\lambda} - \mathbf{j}^T\mathbf{C}\boldsymbol{\lambda},$$

because the journal profile $\mathbf{j}$ is fixed and the factor of $\frac{1}{2}$ does not depend on $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_k\}$, the vector of weights to be found. Here, $\mathbf{C}$ denotes the matrix of community profile vectors stacked side-by-side, $\mathbf{1}$ denotes a vector of ones and $\mathbf{0}$ denotes a vector of zeros.

The R package `scrooge`, available on GitHub[9], provides utilities to calculate these points.

[9] `https://github.com/Selbosh/scrooge`

In the case of singleton community, such as *Stata Journal*, $\boldsymbol{\lambda}$ is simply an indicator vector for its community ID. This is because $\mathbf{C}\boldsymbol{\lambda} = \mathbf{j}$, i.e. the journal profile itself lies inside the convex hull of community profiles. The point $\mathbf{C}\boldsymbol{\lambda}$ is usually, but not necessarily, near to the profile of the community to which the journal was assigned by the community detection algorithm.

For example, consider the grouping of statistics journals suggested earlier by the Louvain method (Table 4.4). The nearest convex combination of communities to the journal *Biometrika* is

$$\boldsymbol{\lambda}_{\text{Bka}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0.13 & 0 & 0.87 \end{pmatrix},$$

suggesting that this journal's behaviour is best described as a mix of communities 5 and 7, rather than just community 7, the one assigned by Louvain. That is to say, *Biometrika* doesn't behave exactly like the average 'statistical theory' journal, but chooses to cite sources a bit like a 'biostatistics' journal as well. The Euclidean distance between this journal profile and the convex community hull is $0.01 > 0$, so some behaviour is still left unexplained by the community structure.

Some journal profiles are much closer to one vertex of the convex hull of community profiles than to any other vertices. The *Journal of the Royal Statistical Society: Series B* has solution

$$\boldsymbol{\lambda}_{\text{JRSS-B}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0.01 & 0.99 \end{pmatrix},$$

while *Statistics in Medicine* has

$$\boldsymbol{\lambda}_{\text{StMed}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

with Euclidean distances 0.02 and 0.03, respectively. Being nearest to a vertex (single community) rather than an edge (mix of communities) of the community hull does not necessarily mean that community describes the journal's behaviour particularly well—*Statistics in Medicine*'s profile is in fact further away than that of *Biometrika*.

Figure 4.6 shows the distances of all the statistics journals from the convex hull of community profiles. The *Journal of Statistical Software* and *Journal of the Royal Statistical Society: Series A* seem to have citation profiles least well described by the Louvain communities.

Figure 4.6: Distances of statistics journal citation profiles from the convex hull of community profiles given by the Louvain method

4.3.3 *Residual analysis*

The analysis of these points and distances may be interesting, but it is still not immediately clear how to determine whether a given clustering explains a citation network well or not. For that, we borrow a standard technique from generalised linear models—residual analysis.

Each journal's outgoing citations can be predicted from the nearest convex combination of community profiles, multiplied by that journal's total observed outgoing citations. That is,

$$\mathbb{E}[\mathbf{X}_j] = \sum_i x_{ij} \mathbf{C}\boldsymbol{\lambda}_j,$$

which may be interpreted as an *offset term* for an implicit Poisson regression with no coefficients, i.e.

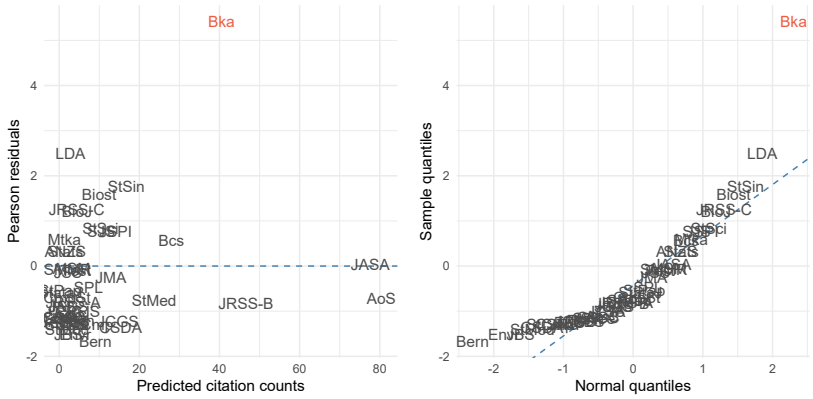
$$\log \mathbb{E}[\mathbf{X}_j] = \log \left| \sum_i x_{ij} \mathbf{C}\boldsymbol{\lambda}_j \right|. \tag{4.11}$$

From these predictions, we obtain errors, called *profile residuals*. As the predictions are counts, we might take the profile residuals to distributed like in a Poisson regression model—with variance equal to the expectation. In which case, profile residuals are defined

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}},$$

where y_i is an observed citation count and $\hat{y}_i$ is the count predicted from the community profiles.

These residuals can be analysed visually, just like those from a conventional generalised linear model, as shown in Figure 4.7.



(a) Residuals versus fitted values

(b) Quantile–quantile plot

Figure 4.7: Profile residual diagnostic plots for Biometrika

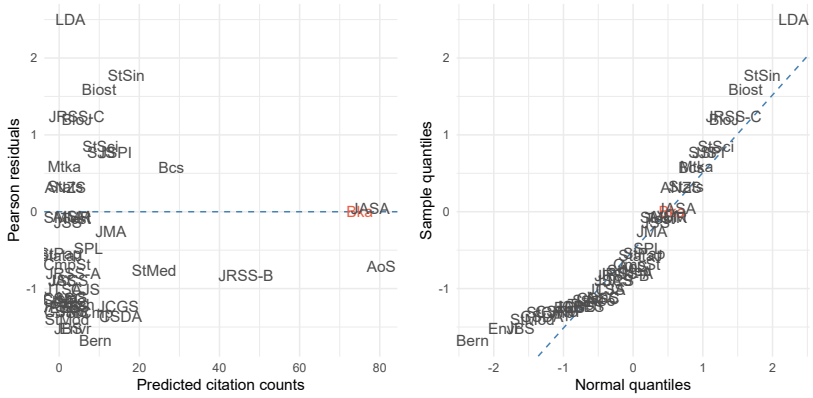
From these diagnostic plots, it appears that *Biometrika* has extremely high profile residuals for citations of itself (highlighted in red). If we conduct the same procedure for almost any other journal, we will find the same phenomenon. Unsurprisingly, an academic journal receives more citations from itself than from an average member of its assigned community (unless, of course, it is a singleton). This can skew a residual analysis by making a journal seem like a poor fit for its community simply because it has a particular propensity for self-citation.

We can control for this by adding a term to the equation (4.11) to build an explicit Poisson regression model for 'excess self-citation'. It takes the form

$$\log \mathbb{E}[\mathbf{X}_j] = \log\left|\sum_i x_{ij}\mathbf{C}\boldsymbol{\lambda}_j\right| + \beta \mathbf{e}_j, \tag{4.12}$$

where β is a coefficient measuring excess self-citation and $\mathbf{e}_j$ denotes a vector with a one in position j and zeros everywhere else.

We obtain $\hat{\beta} = 0.617$ and s.e.$(\hat{\beta}) = 0.115$. Residual diagnostic plots for the updated model are given in Figure 4.8. Now citations from *Biometrika* to itself have a profile residual of zero.



(a) Residuals versus fitted values

(b) Quantile–quantile plot

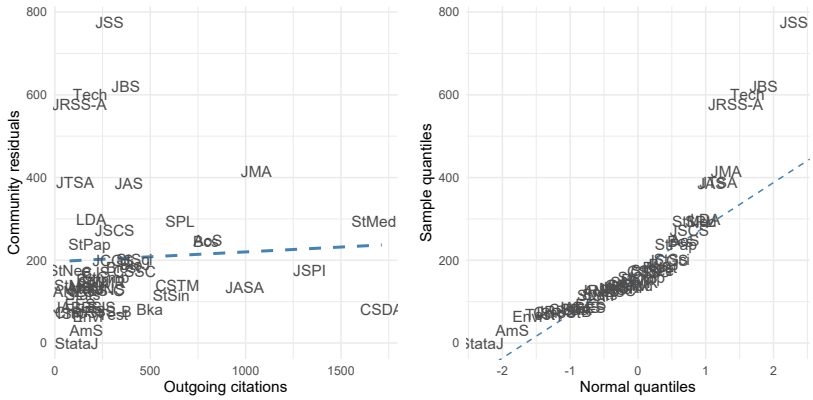
Figure 4.8: Profile residual plots for Biometrika, accounting for excess self-citation

Researchers may not have the patience to analyse the profile residuals for every journal—that would amount to inspecting 47 pairs of plots for our statistical journals dataset. Rather, analysis of profile residuals is a secondary step, following diagnosis of *community residuals* for the whole network.

Community residuals are simply the sum of squares of the profile residuals. Thus, each journal has a single community residual and all community residuals can be viewed in a single plot. Assuming the normal approximation to the Poisson holds, then community residuals should have approximately a χ^2-distribution with $n-1$ degrees of freedom, where n is the number of journals/nodes in the network.

Community residual plots are given in Figure 4.9. There appear to be some journals with extremely high profile residual sums of squares. As seen above, part of this can be explained by the model failing to account for excess self-citation.

The effect of adding a self-citation term to the model is clearly illustrated in Figure 4.10: the community residuals for the 'outliers' are considerably reduced. This implies that some of the discrepancy between community behaviour and individual journal behaviour can be attributed to excess self-citation. Moreover, whilst we do not necessarily accuse any of the journals here of deliberately citing themselves to boost their Impact Factor, if such manipulation were taking place then community residual analysis might be a good way of detecting it.



(a) Residuals against fitted values

(b) Quantile–quantile plot

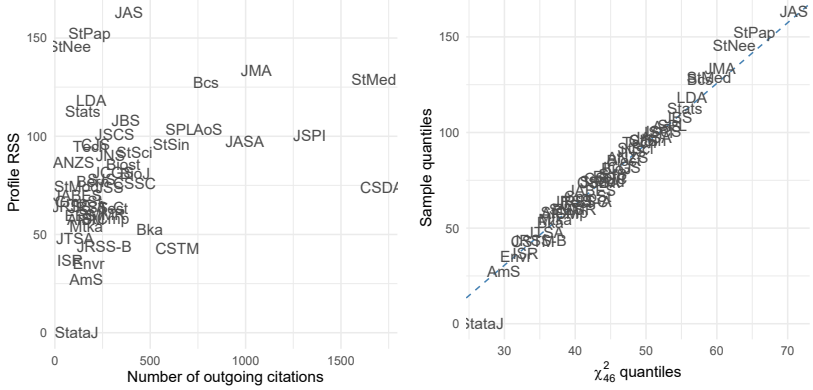
Figure 4.9: Community residual plots for the Louvain method

(a) Uncontrolled against controlled, with line of $y = x$

(b) Red = uncontrolled; blue = with self-citation term

Figure 4.10: Comparisons of community residuals from a 'null' model against those from a model controlling for self-citations

Standard community residual plots, controlling for self-citation, are given in Figure 4.11. Comparing this with Figure 4.9, it is clear to see that, for this example at least, the model assumptions only hold if excess self-citation is accounted for.



(a) Residuals against bibliography length (b) Quantile–quantile plot

Figure 4.11: Community residual plots, controlling for self-citation

Community residuals appear to be χ^2-distributed[10] and mostly uncorrelated with bibliography length. This implies there is no easily-detectable community structure in the data not already explained by the output of the Louvain method (Table 4.4).

Further exploratory data analysis is possible: if other covariates were available, such as the number of articles published in each journal, number of authors per paper or other similar metadata, then we could plot the community residuals against these to investigate any other potential sources of structure.

[10] The χ^2_{46}-distribution has so many degrees of freedom that it is approximately normal. We can plot community residuals against normal distribution quantiles instead and the graph looks almost the same.

4.4 *Concluding remarks*

Community detection is a difficult problem. Though the last decade or so has seen a number of new community detection methods proposed, with ever improving scalability, the problem remains unsolved, especially for weighted, directed networks.

As algorithms and metrics proliferate and data sets expand in size, it is increasingly difficult to compare the quality of outputs from different procedures. Expert judgment is neither scalable nor reproducible, and in bibliometrics, some so-called 'accepted' classification schemes do not necessarily reflect the data.

Community detection problems often have no 'ground truth', but it is nonetheless possible to diagnose when algorithms fail to detect structure that is clearly present. Moreover, a thorough analysis should not simply pick whichever algorithm attains the highest score. As well as trying to understand the differences in how groupings describe behaviour, we can also investigate the varying strengths of communities within a single algorithm's output.

Profile residuals, community residuals and diagnostic plots, introduced in the last section, provide a familiar statistical framework for identifying outliers, uncaptured structure and anomalous be-

haviour, including excess self-citation in communities. Using these techniques should—at least for citation networks and other graphs where we are primarily interested in out-degree distributions—allow us to analyse groupings (whether proposed by experts or as the output of algorithms) in a structured way.

Calculating these quantities is a combination of standard techniques: quadratic programming, generalised linear models and common data visualisations. Convenience functions to perform these tasks are available in the `scrooge` package, which is currently in development.

Further work

This chapter has demonstrated a basic residual analysis for the Louvain method applied to the dataset of 47 journals first studied by Varin et al. (2016). We saw in Table 4.7 that, according to modularity—one flawed measure of community quality—the Louvain method (along with another modularity optimisation algorithm) yields the 'best' grouping, with hierarchical clustering in second place.

A follow-up analysis might look at the profile residuals and community residuals for the communities detected by each method, to see if this is 'really' the case. Moreover, the granularity offered by the log-linear model (4.12) lets us explore other aspects of the data, such as which journals have the greatest propensity for self-citation (measured by the coefficient β) and—with additional data—whether this behaviour changes over time.

Modelling categorical terms, or at least summarising and visualising them, would allow us to see which communities (or publishers, countries etc.) are 'strongest', measured by homogeneity of their members' behaviour. Visualising by category within a grouping may help identify communities that are unusually strongly linked and should be merged.

Examples in the previous section used Pearson residuals, but deviance residuals could also be used. From these we could calculate residual deviance, and compare this statistic with other measures, such as modularity, as a quantitative score for community detection.

5
Citation data and where to find them

5.1 Seeking citation data

In earlier chapters, analyses were performed and models fitted on citation data from the Clarivate Analytics (formerly Thomson Reuters; ISI) Web of Science, which is a commercial database to which the public access is restricted. Clarivate and Thomson Reuters were generous enough to provide us with citation data for academic study, but we cannot assume that such access will be granted to others in the future.

All scientific analyses should be reproducible, and new statistical methodologies are of dubious worth if there are no data available on which to apply them. Where, then, can the reader obtain new citation data to verify and reapply the techniques described in this thesis?

In this chapter, we demonstrate several approaches for extracting pairwise citation data from online databases, some open and some slightly less open, but all free (in terms of cost) to access for academics. Many reviews in bibliometric literature of citation databases are concerned with paper- or author-level citation counts, or the number of documents that are indexed in a subject area, as a measure of 'coverage'. However, few if any articles describe reproducible methods of constructing a cross-citation table that represents a citation network of journals, authors or institutions in a particular time window.

Varin et al. (2016) studied a matrix of citations between 47 statistics journals in 2001–2010. This exact data set is provided, for convenience, in the **scrooge** package. Originally, it was extracted from the Web of Science database. Suppose that we wish to produce an updated version of this matrix with more recent data, or a modified one for a different subject area or specialism, with more journals added.

In the following sections we will briefly view the different sources from which one might be able to obtain citation data for analysis: the Web of Science's Journal Citation Reports, Elsevier's Scopus database, the Google Scholar search engine, Microsoft Academic and its APIs, and finally the new Open Citations Corpus in conjunction with CrossRef.

Using several of these resources, we show, with reproducible R code (R Core Team, 2019), how to build a 4×4 citation matrix for four prominent statistics journals (and the obstacles that researchers may face in the process). This size of data set is small enough to be reproduced quickly without excessive demand on network bandwidth, computation time or API rate limits, but large enough to fit a Stigler export scores model (Stigler, 1994; Varin et al., 2016). The procedure can then be easily extended to more or different journals by swapping in the corresponding journal names or ISSNs, and may be adapted without too much difficulty to citations aggregated by author or institutional affiliation rather than by journal.

Though some tools exist for wrangling bibliometric data from different sources, for example R package **bibliometrix** (Aria and Cuccurullo, 2017), they often assume the user already has the data saved to disk. This chapter on the other hand describes a full pipeline from source to analysis.

After demonstrating the different interfaces we will give some examples of analyses recently made much easier with these tools, and discuss the relative convenience and ease of access of the tools available.

5.2 *Web of Science*

Perhaps the most well-known resource for citation information is the Web of Science, by Clarivate Analytics, publishers of the notorious journal impact factor metrics. This database—and indeed the impact factor—has a long history, pre-dating the information age (Garfield, 2006), but the Internet has spawned several modern competitors.

Mongeon and Paul-Hus (2015) suggested that journal coverage in the Web of Science, as the name might suggest, tends to favour natural sciences and engineering over arts, humanities and social sciences. Moreover, they found that the database overrepresents English-language publications.

Access to the Web of Science is by institutional subscription and provides metrics such as total citation counts, Impact Factors, Eigenfactors and Article Influence Scores, displayed on the Journal Citation Reports (JCR)[1] online dashboard.

[1] `https://jcr.clarivate.com`

5.2.1 *Downloading JCR data*

In addition to these summary measures, the JCR interface visualises journal relationships in the form of tables and charts representing annual citation counts between journals, under the 'Citing Journal Data' and 'Cited Journal Data' tabs. These datasets can be downloaded to disk if one agrees to the company's terms of use. However, there is no API to access many journals at once, and downloading 'excessive amounts of Content' is forbidden by the terms of use.

Interested parties can choose to purchase the data, and there is an exception for researchers in bibliometrics, who can use the resource but only by formal written request. Thomson Reuters (Clarivate) have previously been kind enough to provide us with batches of citation data in a convenient format. However, we—and

others who might hope to replicate our analyses—may not necessarily be able to rely on the company acceding to such requests in the future. Moreover, the data received may not be the same as the data actually used to calculate the published impact factor, as Rossner et al. (2007) described as they unsuccessfully attempted to replicate published impact factors from purchased data.

Due to these obstacles, the remainder of this chapter focusses on alternative data sources.

5.3 Scopus

Elsevier's Scopus is a competing citation database to Clarivate's Web of Science. The rival data set has a suite of rival metrics to boot, including CiteScore, which is similar to Journal Impact Factor, and Scimago Journal Rank (SJR), comparable to Eigenfactor. Though Elsevier charges for access to the Scopus database itself, their journal rankings and metrics are freely available to non-subscribers. Their application programming interface (API) may be used by academics so long as the usage complies with certain terms[2], including: "Public sharing of data for purpose of reproducibility with a specific party is permissible upon written request and explicit written approval." This data-sharing condition, generating the familiar "data are available upon request" footnote in publications, introduces a barrier to efficient replication and scrutiny of results.

[2] https://dev.elsevier.com/academic_research_scopus.html

Shortly after its release in 2004, Neuhaus and Daniel (2006) made an early comparison of Scopus with the Web of Science, finding that it covered more journals, but over a shorter time, and offered a complement rather than necessarily a substitute to Thomson Reuters' offering. More recently, Martín-Martín et al. (2018) compared Scopus's coverage of citation counts, relative to the Web of Science and Google Scholar. Over many academic subject areas, they found strong correlations in citation counts between the three repositories, and that Google Scholar covers a 'superset' of publications found in the two more traditional databases, including more citations from theses, books, conference proceedings, preprints, working papers and other non-journal formats. However, Scopus and the Web of Science provide better structured metadata. Mongeon and Paul-Hus (2015) found that Scopus carries similar biases to the Web of Science in its over-representation of English-language, scientific publications rather than arts and humanities and texts in other languages. Neither of the two repositories has comprehensive coverage, but they are complementary when it comes to different publication types, such as journal articles versus books.

5.3.1 Citation data via the Scopus Search API

For programmatic access to Scopus, anyone—who agrees to certain terms—can create an API key on the Elsevier Developers web

site[3]. This grants the ability to send HTTP requests and make basic queries in Scopus Search, such as looking up articles by DOI or key word, or retrieving journal metrics by ISSN. With the aid of the package **rscopus** (Muschelli, 2019) we can construct the appropriate queries from R. Elsevier provides some tips on getting started[4] online. An example request might be

[3] https://dev.elsevier.com/apikey/create

[4] https://dev.elsevier.com/tips/ScienceDirectSearchTips.htm

```
library(rscopus)
result <- scopus_search("srctitle(biometrika) AND pubyear IS 2019", count = 5, max_count = 5)
```

which returns metadata for five articles published last year in *Biometrika*, including their titles and DOIs:

1. Conjugate Bayes for probit regression via unified skew-normal distributions (DOI: `10.1093/biomet/asz034`)
2. Fast exact conformalization of the lasso using piecewise linear homotopy (DOI: `10.1093/biomet/asz046`)
3. Distributional consistency of the lasso by perturbation bootstrap (DOI: `10.1093/biomet/asz029`)
4. Sequentially additive nonignorable missing data modelling using auxiliary marginal information (DOI: `10.1093/biomet/asz054`)
5. Bayesian jackknife empirical likelihood (DOI: `10.1093/biomet/asz031`)

In order to get 'cited by' data from articles in Scopus, it is necessary to e-mail their 'integration support' team to authenticate your API key, as access to this field is switched off by default. (We are grateful to Dave Santucci from Elsevier for granting us access to this API.)

As with other repositories, it is possible to look up articles by journal ISSN and by publication date. The Scopus `refeid` field records the `eid` (an article identifier) of each article that cites the current one. For example, the article Varin et al. (2016) has a Scopus `eid` of `2-s2.0-84955176281`. We can look up all articles that have cited it as follows:

```
library(rscopus)
query <- "refeid(2-s2.0-84955176281)"
result <- scopus_search(query)
gen_entries_to_df(result$entries)$df
```

The resulting data frame is presented in Table 5.1. It is easy to see that we have all information necessary to determine the containing journal of the citing articles—at least, those that the API tells us about. (In fact, much more article metadata is returned than shown here, but has been omitted to save space.)

With this demonstrated, it is straightforward to extend to many articles and their respective citations. Hence, we search for all the articles in our journals of interest, then iterate over them to list all the citations, filter the journals we care about and count them up in a table. Example code follows.

```
library(rscopus)
# All articles in journals and time window of interest
```

DOI	Date	$\ldots$
10.1002/jtr.2316	1 January 2020	...
10.1214/18-STS686	1 May 2019	...
10.1109/ACCESS.2019.2937220	2019	...
10.1007/s10489-016-0861-4	1 August 2018	...
10.1016/j.joi.2018.06.010	August 2018	...
10.1080/00031305.2017.1360794	3 July 2018	...
10.1098/rsos.171085	December 2017	...
10.1111/rssb.12233	November 2017	...
10.1111/obes.12185	October 2017	...
10.1007/s11192-017-2471-2	1 October 2017	...
10.1007/s11205-016-1407-1	1 September 2017	...
10.1007/s11943-017-0201-0	1 April 2017	...
10.1214/16-AOAS896G	December 2016	...
10.1108/MF-12-2014-0315	11 April 2016	...
10.1371/journal.pone.0143460	1 December 2015	...

Table 5.1: Articles citing Varin et al. (2016), according to Scopus

```
query <- "issn(0090-5364 OR 0006-3444 OR 1369-7412 OR 0162-1459)
          AND pubyear > 2009"
articles <- scopus_search(query)$entries %>% gen_entries_to_df %>% .$df

citations <- rowwise(articles) %>% select(cited_journal = `prism:publicationName`,
    eid) %>% # For each article, list the articles that cite it
mutate(refeid = list(sprintf("refeid(%s)", eid) %>% scopus_search() %>% .$entries %>%
    gen_entries_to_df() %>% .$df %>% count(citing_journal = `prism:publicationName`))) %>%
    ungroup() %>% tidyr::unnest(refeid)

# Filter and cross-tabulate citations
citations %>% filter(citing_journal %in% cited_journal) %>% count(cited_journal,
    citing_journal, wt = n) %>% xtabs(n ~ citing_journal + cited_journal, data = .)
```

See the resulting counts in Table 5.2. The table is broadly similar to other database results, as we shall see in later sections.

	Annals of Stats	Biometrika	JASA	JRSS-B
Annals of Statistics	2213	240	481	287
Biometrika	368	356	355	205
JASA	915	473	1250	450
JRSS-B	369	167	324	248

Table 5.2: Citation flow from statistics journals in 2010–20 (rows) to the same journals in those years (columns), according to Scopus

Alternatively, one can count citations in the opposite direction, via the Scopus Abstract Retrieval API and its 'REF' view, which returns a list of an article's outgoing citations. As with the 'cited by' data discussed above, this field is disabled in the API by default and must be explicitly authenticated. We did not request such permissions so have not tested this approach.

5.4 Google Scholar

In the bibliometrics literature, Google Scholar is often mooted as an alternative citation resource, but such utility in our view is limited. Whilst useful as a search engine, it has no API other than its web interface, making it inefficient for automated data retrieval. The only way to use it programmatically is to make repeated search queries and then scrape the results pages—a tedious process.

The R package **scholar** (Keirstead, 2016) offers functionality for this purpose, but it can only yield summary statistics already reported by the Google web pages, such as the number of articles that an author has published or the number of unique containing journals, author-level citation metrics such as *h*-index, and simple construction of co-authorship networks.

In theory, one might be able to query several articles, look up the articles which have cited them, and then derive some sort of journal-to-journal or author-to-author citation network, but in practice this would be too bandwidth-intensive with the inordinate number of necessary page requests, preventing any kind of scalability, as well as most likely violating Google's terms of use. Alternatively one can do it by hand, which is extremely laborious: see for example Prins et al. (2016), who reportedly spent dozens of hours on such a task.

For these reasons we do not consider Google Scholar a feasible source for the pairwise citation data we seek, unless the company changes tack and decides in the future to release a public API.

5.5 Microsoft Academic

Microsoft Academic *Search* was an experimental service that ran from 2009 until 2012. It was originally pitched as a rival to Google Scholar (as well as older services such as *Scopus* and the *Web of Science*), but was paid little interest by bibliometricians, covering a smaller proportion of publications and citations than competing databases, and was not updated at all after 2012 (Van Noorden, 2014; Harzing, 2016). Relaunched in 2015, Microsoft Academic[5] is built on top of results from the Bing search engine (Knies, 2014), using the Microsoft Academic Graph, which models 'real-life academic communication activities' via six types of entities: field of study, author, institutional affiliation, paper, journal/conference series and event (Sinha et al., 2015).

[5] https://academic.microsoft.com

Preliminary studies (Harzing, 2016) suggested that Microsoft Academic's coverage of the literature outperformed the *Web of Science*, and was competitive with *Scopus*, though did not index as many citations as Google Scholar, from the outset. This has improved with time (Harzing and Alakangas, 2017). Hug et al. (2017) explored the feasibility of Microsoft Academic as a tool for bibliometric analysis, finding the application programming interface (API) makes automated data retrieval and processing much eas-

ier than does Google Scholar (which does not support automated queries) whilst also offering more structured metadata, albeit not as rich as that found in Scopus or the Web of Science.

Thelwall (2017a; 2017b; 2018a; 2018b; Kousha et al., 2018) evaluated Microsoft Academic's accuracy at retrieving specific journal articles, finding a high rate of precision and recall, and strong correlations with *Scopus* for citation counts, detecting slightly more citations and many more 'early' or in-press citations. Thelwall (2018a) points out, however, that Microsoft Academic is not suitable for formal research evaluations, like the UK REF, because it is easy to manipulate results by spamming fictitious or low-quality documents to the Web (this has been demonstrated on Google Scholar; see López-Cózar et al., 2013).

Many of the aforementioned studies comparing 'coverage' of the different citation repositories were performed by querying a topic or institution known to the authors, so as to provide a human-interpretable baseline for the results. Though of course pragmatic, for want of any actual 'ground truth' citation database, this method inevitably introduces a selection bias, so some of the conclusions about which repository offers the best 'coverage' may not hold for other fields, publishers, languages and so on.

5.5.1 Counting citations with the Microsoft Academic API

In this sub-section we use the R package **microdemic** (Chamberlain, 2018), a wrapper around the Microsoft Academic API (Sinha et al., 2015) that simplifies the process of building queries and parsing the responses into a useful format. Alternatively, one can make HTTP requests by hand, using the documentation on the Microsoft Azure web site[6].

[6] `https://docs.microsoft.com/en-us/azure/cognitive-services/academic-knowledge/queryexpressionsyntax`

The API has three main branches, one of them being 'Interpret', designed for providing natural-language queries and autocompletion in search boxes. This might be handy for a future user-facing application, but we focus our attention for now on the other endpoints: Evaluate, which queries the Microsoft Academic database and returns matching results; and CalcHistogram, which provides corresponding metadata, such as the number of matching results to a query, broken down by year, institution and so on.

To obtain citation counts, we can use a workflow something like the following.

Firstly, obtain an API key from Microsoft. Unlike the Scopus API, the process of requesting an authenticated key is entirely via a web form. API keys are free for academic use of Microsoft Academic, and indeed any use: at the time of writing there is no paid tier of the service, so the only limitation is staying within appropriate rate limits: up to 10,000 transactions per day and one Evaluate query per second.

Let's query the 'Evaluate' API for entities with the journal name *Biometrika* or *Annals of Statistics*:

```
library(microdemic)
ma_evaluate("Or(JN = 'biometrika', JN = 'annals of statistics')")
```

Id	DJN	CC	PC
172180718	Biometrika	505908	7267
119757635	Annals of Statistics	438513	5470

Table 5.3: Example response from a Microsoft Academic 'Evaluate' API query

The response, shown in Table 5.3, seems slightly arcane, but the key columns of interest to bibliometricians are `Id`, a unique identifier for the journal; `DJN`, the 'display' journal name; `CC`, the total citations received and `PC`, the total number of publications by that journal.

To look up articles, we query 'paper' entities with respective journal identifiers. The identifier of the journal containing the paper is a `Composite` attribute so we use syntax as follows. Paper entities have a lot of fields, but useful ones include the article `Id` or `DOI`, the journal title `VFN` ('volume full name'), the article title `DN` ('display name'), year of publication `Y` and the references list `RId`.

```
ma_evaluate("Composite(Or(J.JId = 172180718, J.JId = 119757635))", atts = c("Id",
    "VFN", "DN", "Y", "RId"))
```

A segment of the response is given in Table 5.4. The reference list or `RId` (not shown) for each article is simply a vector of article identifiers, e.g. (1554944419, 2110065044, 2135046866, 1988790447, 2912934387, 2982720039, ...).

VFN	DN
Annals of Statistics	Least angle regression...
Annals of Statistics	Greedy function approximation:...
Biometrika	Ideal spatial adaptation by wa...
Annals of Statistics	Additive logistic regression :...
Annals of Statistics	The Dantzig selector: Statisti...
Annals of Statistics	Boosting the margin: a new exp...
Annals of Statistics	Simultaneous analysis of Lasso...
Biometrika	Longitudinal data analysis usi...
Annals of Statistics	High-dimensional graphs and va...

Table 5.4: Example response for a Microsoft Academic query with composite attributes

Equipped with these tools, we can now produce a citation matrix. As we are limited to 1000 records per query, it is important to check how many records there are in total. The CalcHistogram API is useful for this. The following query returns the output given in Table 5.5.

```
ma_calchist("And(Composite(Or(J.JId=62401924, J.JId=172180718,
                              J.JId=119757635, J.JId=145009937)),
                Y=[2010, 2020])",
    atts = c("Id", "DOI", "DN", "J.JN"))
```

From this, we see that the four statistics journals together published 3834 articles since 2010. (Why it says there were $64 = 4^3$

attribute	distinct_values	total_count
Id	3834	3834
J.JN	64	3834
DN	3834	3834
DOI	3768	3769

Table 5.5: Example response from a Microsoft Academic 'CalcHistogram' API query

distinct values for journal name `J.JN` is not clear but it does this for all such queries; the actual corresponding `histogram` in the results correctly shows four names.)

Passing the same query now into the Evaluate API returns the 3834 articles and their reference lists. Merging the table with itself by cited article ID produces a data structure of citing journals and cited journals, which can be aggregated into the contingency table 5.6.

	Annals of Stats	Biometrika	JASA	JRSS-B
Annals of Statistics	1056	117	214	100
Biometrika	470	427	399	237
JASA	1033	488	1297	458
JRSS-B	456	184	351	232

Table 5.6: Citation flow from statistics journals in 2010–20 (rows) to the same journals in those years (columns), according to Microsoft Academic

Compare this 4×4 citation matrix with that generated from Scopus in Table 5.2. The pattern of citation flows is similar. For example, *Biometrika* cites *Annals*, *JASA* and itself about the same amount, but *JRSS-B* half as often; *JRSS-B* cites, in descending order, *Annals*, *JASA*, itself and *Biometrika*; and *JASA* has a considerably longer reference list (more outgoing citations) than the others in the same period. However, *Annals of Statistics* gives out around twice as many citations according to Scopus than according to Microsoft Academic. This could be caused by inclusion of certain article types that do not appear elsewhere.

5.6 Open Citations Corpus

Just as data scientists nowadays may tend to opt for open-source tools such as R, Python, LaTeX and JavaScript over proprietary ones such as S-Plus, Office and Flash, an open-access data movement is gathering pace.

Shotton (2013) described the lack of openly accessible citation data as a 'scandal', with universities paying thousands of pounds every year for access to the Scopus or Web of Science citation repositories. Since 2010, Shotton has led a project called the Open Citations Corpus, which leverages CrossRef—the organisation that provides document object identifiers (DOIs) to academic works—to form a semantic graph of citations amongst open-access journals, pre-print repositories and (cooperating) traditional publishers.

5.6.1 *Open citations via SPARQL*

An example query follows, using the SPARQL language. This requests the title, publication date and DOI of all articles published in the journal *Biometrika* in 2012.

```
PREFIX cito: <http://purl.org/spar/cito/>
  PREFIX dcterms: <http://purl.org/dc/terms/>
  PREFIX datacite: <http://purl.org/spar/datacite/>
  PREFIX literal: <http://www.essepuntato.it/2010/06/literalreification/>
  PREFIX biro: <http://purl.org/spar/biro/>
  PREFIX frbr: <http://purl.org/vocab/frbr/core#>
  PREFIX pubdate: <http://prismstandard.org/namespaces/basic/2.0/publicationDate>
  SELECT ?title ?date ?id ?doi WHERE {
    ?article frbr:partOf ?issue ;
    pubdate: '2012'^^xsd:gYear .
    ?issue frbr:partOf ?volume .
    ?volume frbr:partOf <https://w3id.org/oc/corpus/br/18712>
      OPTIONAL {
        ?article dcterms:title ?title ;
        pubdate: ?date ;
        datacite:hasIdentifier ?id .
        ?id literal:hasLiteralValue ?doi ;
        datacite:usesIdentifierScheme datacite:doi
      }
  }
LIMIT 100
```

This query, which has a rather cumbersome syntax[7], returns 11 results; this is less than one might expect to see for a whole year's worth of publications. For other publication years, too, we only retrieve around 10–11 articles each time, so the corpus or its corresponding metadata seem incomplete. (For reference, a quick perusal of the publisher's web site reveals that *Biometrika* actually publishes some 80 articles each year.)

[7] 'An article, which is part of an issue, which is part of a volume, which is part of the journal, which has the title *Biometrika*'

5.6.2 *Open citations via R*

As an alternative, we can retrieve DOIs instead directly from CrossRef, with a bit more success. For that, we first use the R package **rcrossref** (Chamberlain et al., 2019) to get the DOIs, then **citecorp** (Chamberlain, 2019) to look up corresponding citations in the Open Citations Corpus.

The procedure is roughly as follows.

1. Find the DOIs for all works published in the journals of interest, using the CrossRef API.
2. Get a list of all the references contained in those articles, via the Open Citations Corpus.
3. Filter the list of cited articles based on whether they appear in the journals and time window of interest.

Researchers interested in article *content*, rather than metadata, can use R packages **crminer**, **fulltext** and **roadoi**, which provide facilities for downloading full text articles freely from open-source repositories; see `https://rOpenSci.org` further information.

4. Aggregate the data, counting the numbers of citations between journals.

And this can be implemented using outline R code of the following form.

```
# Step 1
library(rcrossref)
citing_articles <- cr_journals(
  issn = c('0090-5364', '0006-3444', '1369-7412',  '0162-1459'),
  filter = list(from_pub_date = 2010,
                until_pub_date = 2020),
  works = TRUE)$data
# Step 2
library(citecorp)
references <- lapply(citing_articles$doi, oc_coci_refs)
# Step 3
library(dplyr)
inner_join(references, citing_articles, by = c(cited = 'doi')) %>%
  # Step 4
  inner_join(citing_articles, by = c(citing = 'doi')) %>%
  count(citing_journal, cited_journal)
```

The method yields citations from articles published in journals during a particular time window to articles in the same journals in the same period. The citing and cited journals and time windows need not be the same—we can vary either of them by carefully filtering the `citing_articles` data frame in step 3 or 4.

In this way, we obtain the following cross-citation matrix for the prestigious statistics journals *Annals of Statistics*, *Journal of the Royal Statistical Society: Series B*, *Journal of the American Statistical Association* and *Biometrika*. The results are printed in Table 5.7, which might be considered a small, modern recreation of the matrix presented in Stigler (1994) and reproduced in the R package **BradleyTerry2** (Firth and Turner, 2012).

	Annals of Stats	Biometrika	JASA	JRSS-B
Annals of Statistics	0	0	0	0
Biometrika	268	252	234	129
JASA	737	364	907	308
JRSS-B	322	136	252	194

Table 5.7: Citation flow from statistics journals in 2010–20 (rows) to the same journals in those years (columns), according to the Open Citations Corpus

Citations from *Annals of Statistics* are conspicuous in their absence from the table. Further investigation reveals that this is because the publisher, the Institute of Mathematical Statistics, is not a listed participant in the Initiative for Open Citations, unlike the respective *Biometrika* and Royal Statistical Society journal publishers, Oxford University Press and Wiley (see I4OC, 2020). Citations *to* the *Annals* are available because the containing bibliographies were provided by co-operating publishers. But querying the *Annals*'s

reference list yields null results, using either the Open Citations Corpus or CrossRef APIs (each article has metadata corresponding to its DOI but the bibliography field is simply empty).

To complete the cross-citation matrix, the academic community can lobby the missing publishers to give permission to CrossRef to include their data in the Open Citations Corpus. In the meantime, what options do we have? We could omit the likes of *Annals of Statistics* from our analysis and focus our attention on the data we do have.

Or, if we are keen to include all journals, we can supplement or replace the data with that obtained from Microsoft Academic (for instance) whose web-scraped data do not follow the same rules for inclusion or exclusion of particular publications.

For the remaining citation flows, we can see that the Open Citations Corpus represents a subset of the data accessible via Microsoft Academic. Some of the more obvious anomalies are explainable, such as the exclusion of *Annals of Statistics* and other journals from the database, as just mentioned. Whether Microsoft Academic's other journal–journal citation counts are all greater than those in the Open Citations Corpus because of better coverage or because of double counting—or a mixture of both—is not immediately clear.

5.7 *Building networks of authors or institutions*

Flexible APIs let us tackle different problems using similar methods. Whereas the Web of Science dataset, studied in other chapters, was inherently aggregated at the level of journals in given years, the individual-level article metadata available in the Open Citations Corpus, Microsoft Academic and Scopus may, to varying degrees, allow us to aggregate papers by author or institutional affiliation instead.

These groupings throw up new problems that are not faced when studying journal networks: papers often have multiple authors, whereas they are usually (modulo pre-prints) published in only one journal at a time. Moreover, each author may have multiple affiliations, which change over time. If a paper co-authored by authors A and B cites another paper co-authored by A and B, is this just a self-citation, or should we count a citation from A to B and from B to A as well? Should all authors and all such flows be weighted equally?

Other issues arise, such as how to identify authors uniquely when different people may have the same name, and one person may have multiple names or abbreviations, and names can change. These issues are left as an exercise for other researchers, who will in any case be able to audit any results presented here by re-running the code and making different decisions.

Let's consider another small example: measuring the citation influence—by way of a (naïve) quasi-Stigler model[8]—between university statistics departments. We can do this by aggregating papers

[8] The quasi-Stigler model (Varin et al., 2016) is described in more detail in Chapter 3.

by the declared institutional affiliation of their authors, rather than the journals in which the articles were published.

In the UK, the Research Excellence Framework is concerned with institutions' research in a global context, assigning ratings such as 'recognised internationally', 'internationally excellent' and 'world-leading'. As such, building a local or national network of university departments' citations and attempting to measure their relative influence on one another is measuring something rather different. Any citation-based ranking would therefore need, at the very least, to include institutions from around the world to be even remotely comparable to a national research assessment—and even then, much global influence may not be measured in citations.

In Microsoft Academic, institutional affiliations have unique entity IDs as well as human-readable names. Examples are the 'normalized' names `university of warwick`, `university college london`, `lancaster university` and `london school of economics and political science`, corresponding to the IDs 39555362, 45129253, 67415387 and 909854389. To look up a paper whose author is affiliated with a particular institution, we can refer either to the name or the ID. The following CalcHistogram query counts the number of articles published by authors affiliated with the University of Warwick each year over the past ten years.

```
ma_calchist('And(Composite(AA.AfId=39555362), Y>2009)', atts = 'Y')
```

The resulting 'histogram', of article counts by year, is shown in Table 5.8.

Table 5.8: Number of articles published by authors affiliated with the University of Warwick each year, according to Microsoft Academic

Year	Count
2010	2327
2011	2631
2012	2805
2013	3202
2014	3493
2015	4000
2016	3992
2017	4056
2018	3912
2019	4551

Making a similar query to the Evaluate API, requesting the `RId` fields, we can thus construct a table of citation relationships between several universities of interest. If we fractionally weight citations according to the number of citation relationships between the citing and cited articles' affiliations, we get a citation matrix like that in Table 5.9.

Going a step further, we can fit a quasi-Stigler model (Varin et al., 2016) to the data, giving a relative ranking of institutions according to their propensity to cite or be cited by one another.

A centipede plot of the Stigler-model export scores and their 95% comparison intervals is given in Figure 5.1. The worst relative errors due to the quasi-variance approximation are −3.6% and +9.5%—small enough to permit the use of comparison intervals

Table 5.9: Citations among a group of UK universities in 2010–2020, from rows to columns

	Birmingham	Cambridge	Lancaster	LSE	Oxford	UCL	Warwick
Birmingham	119.3	17.3	1.0	6.0	10.9	9.8	3.9
Cambridge	15.3	310.6	24.7	33.2	29.7	31.9	26.1
Lancaster	1.0	17.4	260.7	4.0	14.3	19.0	40.4
LSE	1.2	18.2	4.0	144.7	22.2	10.8	3.3
Oxford	10.9	34.4	20.5	8.6	230.7	36.2	38.4
UCL	19.4	49.2	29.1	9.6	43.8	329.6	18.5
Warwick	6.0	17.6	53.7	7.6	47.1	26.7	201.3

in the plot. The ranking is not clearly defined, with a large level of uncertainty. Birmingham has an extremely wide comparison interval and an export score near zero, suggesting there is not a great deal of information about this institution from which we can conclude much.

This model is also based on some extremely bold assumptions—including that citations are generated independently, and that researchers consider institutional affiliations when choosing which papers to include in their works' bibliographies. Nonetheless, it is an example of the kind of analyses that *can* be done with new sources of citation data; whether or not they *should* be done is another matter.

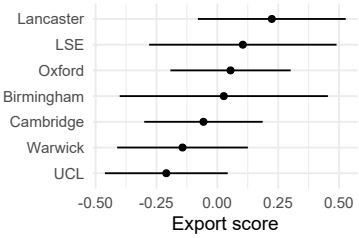


Figure 5.1: Stigler-model export scores and 95% comparison intervals for Microsoft Academic citation data between UK universities in 2010–2020

5.8 *Computing impact factors*

Another interesting exercise is to use Microsoft Academic's CalcHistogram tool to compute our own average per-article citation counts for journals, either to compare with the Clarivate-published journal impact factor (and Scopus CiteScore) or to provide estimates where published metrics are not available. For example, at the time of writing, the latest edition of Journal Citation Reports is 2018. While waiting for the 2019 edition to be published, we might like to compute our own estimates using the latest citation data available.

Furthermore, we might propose that a citation window of two to five years is not enough to encompass the delay between a mathematical sciences paper being published and being recognised and cited by later published works. The flexibility of the CalcHistogram API allows the calculation of citation averages ('impact factors') over 10 years or more.

Table 5.10 compares the latest (2018) JCR impact factors for our four statistics journals with a 10-year metric we calculated ourselves with Microsoft Academic.

Journal	2-year IF	5-year IF	10-year 'IF'
Annals of Statistics	2.901	4.497	37.1
Biometrika	1.641	2.466	17.9
JASA	3.412	3.639	22.8
JRSS-B	3.278	5.327	45.8

Table 5.10: Published impact factors from the 2018 edition of the Journal Citation Reports (Clarivate Analytics, 2019) compared with a 10-year 'impact factor' we have computed from Microsoft Academic data for 2010–2019

From this quick example it is noticeable how sensitive the impact factor appears to be to the time window chosen. Whereas *Journal of the American Statistical Association* has the largest score over two years, over longer periods it is *Series B* that comes top and *JASA* is third. The published five-year metric gives the same ranking as our computed ten-year one.

The following R code demonstrates how one can compute this metric for a single journal such as *Biometrika*.

```
# Make the Microsoft Academic query
result <- ma_calchist("And(Composite(J.JN = 'biometrika'), Y=[2010,2020])",
```

```
                    atts = 'CC', count = 1000)
# Extract the frequency table
freqs <- result$histograms$histogram[[1]]
# Compute the citation average
with(freqs, sum(count * as.integer(value)) / sum(count))
```

5.9 Discussion

In this chapter we demonstrated several methods by which a bibliometrician, or anybody else interested in research assessment, can download pairwise citation data.

These data sources have differing levels of freedom, from open-source (or 'free as in free speech') and no-cost ('free as in free beer') to more restrictive levels of access.

The free and open-source Open Citations Corpus, when used alongside the CrossRef API, in principle provides an easy way to download citation data for our analyses, but in practice the non-cooperation of certain journal publishers leaves gaps in the data and undermines the Corpus's utility. Still, its freedom, 'as in free speech', is the standard to which we must aspire, for reproducibility.

Microsoft Academic is 'free as in free beer' as it currently costs nothing to use and indeed there is no paid tier of the API, though as part of the Azure cloud services ecosystem, in principle the corporation could change the terms of use, discontinue the service or start charging for access in the future. Gaps in Open Citations data effectively make Microsoft Academic, by default, the 'freest' way to access data from the majority of journals, and the API is straightforward to use for this purpose, with the aid of wrapper R packages such as **microdemic** or otherwise. Along with the Open Citations Corpus, Microsoft Academic is the newest source of citation data, and perhaps the least used in the bibliometric literature. Increased usage of this resource could have a profound effect.

Further down the freedom scale we find Scopus, which provides an API for academic use, but requires explicit authentication, and presumably an existing (paid) institutional subscription, such as the University of Warwick's, to access it. Once those prerequisites are in place, however, the Scopus API is about as easy to use as that for Microsoft Academic or the Open Citations Corpus. The Web of Science might find itself near the bottom of the scale: in the past the proprietors have provided batches of data for academic use, but this is not guaranteed in the future and there is no reproducible API; their data has been called 'opaque' (Rossner et al., 2007). Conversely, though Google Scholar is free to access without subscription, it is primarily designed as a search engine. With no public API, obtaining citation data programmatically from Google Scholar is not convenient, making it not a particularly useful resource for this purpose.

What does the future hold? We can expect the Open Citations

Corpus to continue to grow. With this and Microsoft Academic freely available for academic use, and—as demonstrated—fairly easy to use with some rather simple R or HTTP queries, citation analysis is opened up to many more people. The journal impact factor may also face increased competition and scrutiny, as it is easier than ever before for individuals to calculate their own citation metrics and perform their own analyses.

6
Research Excellence Framework & journal rankings

6.1 Introduction

The Research Excellence Framework (REF; successor to the Research Assessment Exercise, or RAE) is the method used by UK funding bodies to evaluate the quality of research. The last REF took place in 2014 and the next one is currently scheduled for 2021. Panels of experts rate universities and research institutions in three categories: impact outside academia, research environment and quality of outputs, based on written submissions. In the sciences and some other fields, submissions are more likely to comprise academic journal articles than books or reports (Wilsdon et al., 2015; Marques et al., 2017).

Expert panels can judge a submission to be 'world-leading' (4*) 'internationally excellent' (3*), 'recognised internationally' (2*), 'recognised nationally' (1*) or unclassified. Results published online[1] describe, for each subject area ('unit of assessment') the proportion of each institution's outputs that were assigned to each of these categories.

[1] https://www.ref.ac.uk/2014

Though it is publicly known which works were submitted for assessment, the ratings are only published in aggregate, by institution and subject: it is not disclosed which rating was assigned to which paper. Thus, it is not obvious what constitutes a '4* paper' or which authors wrote them. However, rumours have long circulated about lists of '4* journals' that peer review panels might use to help them determine the quality of articles (Oswald, 2007).

Given that the purposes of the REF are explicitly '[to] establish reputational yardsticks' (i.e. rank academic departments) and 'to inform the selective allocation of funding for research' (REF web site 2019), it is not surprising that it has had an effect on institutional behaviour, allegedly increasing the number of staff hired on short-term contracts that coincide with the assessment period (Jump, 2013), changing the way departments submit members of staff and publications for evaluation (Marques et al., 2017) and increasing productivity just before the deadline (Groen-Xu et al., 2017).

The popularity of journal-level citation metrics such as the impact factor raises the question: might some expert panels be influenced (consciously or otherwise) by a journal's reputation or

citation count when judging an individual paper?

In this chapter, we investigate the extent to which research institutions' REF ratings (for outputs) might be attributed to the journals in which their outputs were published. Paper-level ratings are missing, but the margins—institutions' REF profiles and the numbers of articles they submitted from each journal—are known, so the research question becomes an 'ecological inference' problem. Using both frequentist and Bayesian approaches, we will estimate latent 'quality' scores for journals, and then quantify the variation in REF results that is explained by these scores. We also compare these scores with published journal citation metrics.

Initially, we demonstrate the methodology on the field of economics, a relatively small and well-defined discipline, which mostly publishes its outputs in academic journals and has a well-established 'Top Five' journals that act as a baseline. Results will then be compared with several other, larger academic fields whose REF-submitted outputs are also mostly in the form of journal articles.

6.2 *Background*

6.2.1 *Modelling research assessments*

Koya and Chowdhury (2017) suggested that there is, for some subject areas, a correlation between journal rankings and REF performance. Their approach computed a 'monetary value' (funding allocation) for each research output as rated in the REF, using a similar method to that described in an earlier blog post by Reed and Kerridge (2017).

Let F be the total amount of funding awarded to an institution based on the REF, let n_3 and n_4 be the number of 3* and 4* outputs and let x_3 and x_4 be the respective monetary value of an output with each rating. According to the then Higher Education Funding Council for England (HEFCE), a 4* output is worth four times as much as a 3* output (Else, 2015), so $x_4 = 4x_3$. The numbers of outputs are known. Then $F = n_3x_3 + n_4x_4 = n_3x_3 + 4n_4x_3$, from which we obtain

$$x_3 = \frac{F}{n_3 + 4n_4},$$

for a given institution and subject area.

Consider the example of general engineering at the University of Cambridge. It was awarded $F = £5{,}328{,}295$ in 2015–16 as a result of its outputs submitted to the 2014 REF[2]. Of the submitted outputs, 37.4% were rated 4* and 55.8% at 3*, for 177.2 full-time equivalent staff. Each staff member was allowed up to four submissions, and funding was allocated assuming that staff submitted this maximum, even if they did not. So the theoretical (not actual) number of outputs was $4 \times 177.2 = 708.8$. Thus $n_3 = 708.8 \times 55.8\% = 395.5104$ and $n_4 = 708.8 \times 37.4\% = 265.0912$, from which we obtain $x_3 = £5328295/1455.875 = £3{,}659.86$ and $x_4 = £14{,}639.43$.

[2] according to the HEFCE 2015–16 funding allocation tables for research

From here, Koya and Chowdhury (2017) studied the relationship between the distribution of an institution's REF scores with the venues in which the outputs were published. REF results do not reveal which article/submission received each rating; the data are only published in aggregate. Koya and Chowdhury (2017) "identified how many of the submitted articles were in top quartile *[sic]* journals" based on impact factors published in the 2013 edition of Thomson Reuters'[3] *Journal Citation Reports*, and compared this proportion with the percentages of articles awarded 4* and 3* ratings[4]. Positive correlations, where found, were weak and only present in some subject areas. Surprisingly, Koya and Chowdhury (2017) did not directly compare the computed 'monetary value' of research outputs with the corresponding bibliometric indicators for each institution.

[3] Now operated by Clarivate Analytics

[4] As pointed out by Hill (2017), this means only a subset of journal articles are being compared with the full range of outputs submitted to the REF—not a like-for-like comparison.

Wilsdon et al. (2015, Section 9.1) commissioned HEFCE to perform a more detailed study of the relationship between bibliometric indicators and REF scores, with privileged access to ratings at the level of the individual outputs. That analysis found low (< 0.5) positive correlations between citation metrics and 4* outputs, but with stronger relationships for some fields such as medicine, biology, chemistry, physics and economics. The strongest predictors were full-text clicks (on Scopus), number of authors, citation count (according to Google Scholar), SJR (a Scopus-published journal metric based on PageRank score), source-normalized impact per paper (a another Scopus metric, similar to a weighted impact factor), tweets, and downloads from the web site *Science Direct*.

In the field of Art and Design, Mansfield (2016) ranked journals according to their popularity in REF submissions, but did not attempt to infer star ratings for the publications.

Stockhammer et al. (2017) investigated the 'grade point average'—the average star rating—of each institution in the 2014 REF, modelling it as a linear function of either the SCImago Journal Rank citation score (SJR) or of journal ratings assigned by the Chartered Association of Business Schools[5]. That analysis, applied to the fields of economics, found a coefficient of determination of up to $R^2 = 89\%$, with the 2014 log-SJR score having a statistically significant effect under their model.

[5] https://charteredabs.org

Italy's research assessment exercise, the *Valutazione Triennale della Ricerca* (triennial research evaluation) began in 2003 with a similar remit to the UK's RAE/REF and was initially 'fully based on peer review'. Franceschet and Costantini (2011) found positive correlations between the peer review assessments and citation metrics, but the strength of the correlation varied between fields, and was particularly weak for journal impact factor. The then-recently proposed h-index (Hirsch, 2005) provided a better approximation.

From 2004, the *Valutazione della Qualità della Ricerca* (research quality evaluation; VQR) introduced a 'dual system of evaluation' using a combination of peer review and bibliometrics. The Italian National Agency for the Evaluation of the University and Research

Systems (ANVUR) compared the results from each approach and found a 'more than adequate concordance', apparently justifying the decision to use bibliometrics. However, this conclusion has been strongly challenged by Baccini and Nicolao (2016), who insist the methodology was 'fatally flawed' and undermines the results for the field of economics and statistics in particular.

Between the RAE2008 and REF2014, Mryglod et al. (2015a) compared departmental h-indices with performance in the RAE, finding a correlation between h-index and certain grade-point averages of RAE results. Using this relationship, they made predictions for the upcoming REF2014 for several institutions and fields. However, in a follow-up after the REF2014 results were published, Mryglod et al. (2015b) reported the predictions "failed to anticipate with any accuracy either overall REF outcomes or movements of individual institutions in the rankings relative to their positions in the previous Research Assessment Exercise". Thus care should be taken in trying to predict one research assessment from the results of another that took place years before.

Our research is not the first attempt at producing a journal ranking from REF results for economics, let alone for academic fields in general. Hole (2017) used a greedy iterative algorithm to assign star ratings to individual papers (assuming these were entirely dependent on the journals in they appeared) minimizing the squared error in predicted ratings for institutions,

$$Q = \sum_{i=1}^{I} \sum_{r=1}^{4} N_i (p_{ir} - \hat{p}_{ir})^2,$$

where N_i is the number of submissions from each institution, p_{ir} is the observed proportion of r-star submissions from that institution and $\hat{p}_{ir}$ is the predicted proportion, based on the imputed ratings. The algorithm first assigns an arbitrary star-rating r to each journal, calculates the objective function Q, then iterates over the list of journals, changing each journal's star rating to that which would decrease Q the most, terminating when a full pass over all journals produces a change in Q smaller than a pre-specified threshold. The analysis of Hole (2017) excluded journals with fewer than five submissions in the REF. Since this would result in the number of submissions no longer adding up to the total number of ratings, they assigned arbitrary ranks to these left-out journals. The results had a correlation of approximately $\rho = 0.5$ with previously-published economics journal rankings.

More recently, Balbuena (2018) adopted a machine learning approach, using a Bayesian additive regression tree model to predict grade point average from a range of institutional covariates, including the number of attributed documents indexed in the *Web of Science* and the proportional intake of students from state schools. However this analysis focussed more on possible inequities in distribution of funding, rather than investigating an explicit journal identity effect.

Yan (2017) used a Metropolis-within-Gibbs sampling regime to fit an ordinal response model to Economics & Econometrics outputs for REF2014. Whilst broadly similar to our approach, their framework is based on a proportional-odds cumulative probit model, which assumes a common set of thresholds between star ratings for all journals. In other words, the increase in difficulty of attaining a 4* rating over a 3* one is the same for every journal. Our analysis fits models for several different subjects and finds that this assumption does not hold, even for Economics & Econometrics.

6.2.2 *Ecological inference*

The previous section provided examples of limited analyses comparing some citation indices and other journal- or institution-level covariates with REF results, and of approaches to produce journal rankings from institutional scores. However, to our knowledge, no principled *statistical* analyses (that is, with quantified uncertainty) of the relationship between journal identities and UK research assessment have been published. Moreover, modelling REF ratings as a function of citation metrics is problematic; criticisms abound of certain indicators under inspection—impact factor and its variants, as well as 'alternative metrics' such as tweets and download counts (e.g. Colquhoun and Plested, 2014; MacRoberts and MacRoberts, 2017). Instead of using a flawed and imprecise proxy such as a citation metric to analyse the relationship between publications and research assessment, one might consider modelling published REF results against the actual journal identities instead. The problem with this approach is that HEFCE (or since April 2018 its successor, Research England) will never publish the individual ratings given to submissions in the REF; indeed they were destroyed upon completion of the research assessment (REF, 2015).

We are therefore left in a quandary: how do we model the effect of journals on star ratings, if we don't know which journal articles received which ratings? What if we wanted to try to infer these publication-level ratings? This would allow us to construct a ranking of *journals*, not just institutions, from the REF results, similar to the work by Hole (2017). Moreover we might attempt to answer the question: is an institution's REF rating simply a function of the journals in which it published? Were that to be the case, it would suggest that the REF is directly measuring prestige rather than quality—a common criticism of citation indices. On the other hand, if an institution's REF score is *more* than the sum of its output journals then it might be used as evidence against using journal-level metrics to assess research quality.

However, as already mentioned, the REF scores are aggregated by institution, not by journal. Journal-level scores must therefore be imputed rather than observed. Such a task—inferring individual-level properties from aggregate data—is known as *ecological inference* or *ecological regression* (Goodman, 1953), typically applied to

estimate voting behaviour in a secret ballot, when exit polls are infeasible or unreliable. Examples include modelling voter transitions between parties (Brown and Payne, 1986) and estimating who voted for the Nazi Party in Weimar Germany (Rosen et al., 2001). A detailed review of the topic is provided by Wakefield (2005). The following is a brief summary.

Sociologists and political scientists often use the term 'ecological inference' to refer to inference on voting populations—for example, voter transitions between elections in a two-party system, or turnout for two demographics.

Consider an election where, to comply with civil rights legislation, authorities in the US desire to compare turnout amongst black and white voters. Suppose for a given electoral district (constituency) i, the demographic makeup is known with proportion X_i of the population black and the remainder white. Overall voter turnout, T_i, is observed for a particular election but the ballot is secret, so turnout among blacks and whites, respectively β_i^b and β_i^w, are unknown. These data yield the following 2×2 table of proportions.

Table 6.1: Observed and unobserved proportions for a two-dimensional voter turnout model

	Vote	Not vote	
Black	β_i^b	$1-\beta_i^b$	X_i
White	β_i^w	$1-\beta_i^w$	$1-X_i$
	T_i	$1-T_i$	

At first glance, it may not appear that one can really glean any information about individuals only from the margins. Via the *method of bounds* however, we can obtain deterministic bounds on (at least one of) the parameters: black turnout β_i^b must be greater than $\frac{T_i-(1-X_i)}{X_i}$ and smaller than $\frac{T_i}{X_i}$, whilst white turnout β_i^w must be between $\frac{T_i-X_i}{1-X_i}$ and $\frac{T_i}{1-X_i}$, to ensure they are valid proportions that add up to one (Duncan and Davis, 1953). For example, if a district's population were 70% black and overall turnout were 40%, then black turnout must be in the range (14%, 57%), but white turnout could still be anywhere in (0, 100%).

Unlike this limited deterministic approach to the ecological inference problem, *ecological regression* or *Goodman regression* (1953; 1959) is one of the first *statistical* solutions. Using the identity

$$T_i = X_i\beta_i^b + (1-X_i)\beta_i^w,$$

one may construct a simple linear regression model of turnout on racial proportions:

$$\mathbb{E}[T_i|X_i] = \alpha + \beta X_i,$$

where $\alpha = \beta_i^w$ and $\beta = \beta_i^b - \beta_w$. A notable criticism is that these voting propensities are assumed to be homogeneous over districts, regardless of the racial mix in each area. Moreover, least squares

does not constrain these parameters to lie within the bounds described above, or even between zero and one (Wakefield, 2005).

Brown and Payne (1986) proposed modelling voter turnout using a convolution of Dirichlet–multinomial distributions, with the response approximated by a multivariate normal distribution. However, this model is sensitive to the choice of prior (Wakefield, 2005). More recently, King (1997) combined the method of bounds with a pseudo-'likelihood' function—equivalent to an asymptotic form of the binomial distribution—and imposed a truncated bivariate normal distribution to tighten the bounds. This approach describes itself as 'a solution to the ecological inference problem', however this claim was criticized as overly optimistic (Cho, 1998; Freedman et al., 1999).

Since then, King et al. (1999) proposed a different solution in the form of an hierarchical Dirichlet–multinomial model where the unobserved probabilities (the voter turnouts by ethnicity) are beta-distributed latent random variables. For a constituency/district i with total voting-age population N_i and observed voter turnout count $Y_i = N_i T_i$ the hierarchical model takes the form

$$Y_i \sim \text{Binomial}(N_i, T_i),$$

where the marginal probability of voter turnout in constituency i is

$$T_i = \sum_{j=1}^{J} x_{ij} \beta_i^j,$$

with x_{ij} denoting the proportion of people of ethnicity j in constituency i, and where the prior constituency-level probabilities of voter turnout, by ethnic group, are

$$\beta_i^j | a_j, b_j \overset{\text{iid}_i}{\sim} \text{Beta}(a_j, b_j)$$

with hyper priors

$$a_j, b_j \overset{\text{iid}_j}{\sim} \text{Exp}(\lambda)$$

and the default hyper-parameter setting $\lambda = 0.5$.

In the two-dimensional (black–white voter turnout) case described above, $J = 2$ and the middle level is

$$T_i = X_i \beta_i^b + (1 - X_i) \beta_i^w.$$

The model generalizes to $J > 2$ ethnic groups (or journals, in our case) and can be further extended to multiple outcomes (beyond binary 'vote or not') by replacing the beta–binomial distribution pair with a Dirichlet–multinomial (Rosen et al., 2001).

In our view, the top level of King's hierarchical model possibly adds an unnecessary random component, for the total turnout should simply be a deterministic, weighted sum of the turnout among each ethnic group. The election result is not an approximation of the counted votes: it *is* the counted votes. All that is necessary is for the βs to be constrained so that the sum over ethnic

groups of voters adds up to the observed overall turnout. This is perhaps more easily said than done, however.

More recent approaches to ecological inference make use of *distribution regression*, by treating the makeup of each electoral district as a probability distribution (Flaxman et al., 2015; Szabó et al., 2016). The basic idea is to project the distributions into a feature space, then fit a regularized regression model, such as kernel ridge regression, using this embedding. Flaxman et al. (2015) used this technique to combine demographic and spatial information and infer the groups who voted for Barack Obama in the 2012 US presidential elections, and again for the 2016 elections (Flaxman et al., 2016).

Rosenman and Viswanathan (2018) derived a 'heteroscedastic Gaussian' approximation to the Poisson binomial log-likelihood, via a central limit theorem, and later applied this to a large voter transition model (Rosenman, 2019), which they term a *Poisson binomial generalized linear model*. Unlike the presidential election studies by Flaxman et al., which used Bayesian techniques, the Poisson binomial GLM is 'purely frequentist'. This offers the advantages of 'simpler fitting procedures, straightforward estimation of individual-level probabilities, and greater model interpretability' at the expense of reduced flexibility (Rosenman, 2019).

6.3 Model

The REF ratings received by institutions on their outputs could be assumed to be drawn from a Poisson binomial distribution (Poisson, 1837), which describes the probability of obtaining K successes in n independent but non-identically distributed Bernoulli trials, with probability mass function

$$\Pr(K = k) = \sum_{A \in F_k} \prod_{i \in A} \pi_i \prod_{j \in A^c} (1 - \pi_j),$$

where F_k is the set of all subsets of k integers that can be selected from $\{1, 2, \ldots, n\}$, for n the number of Bernoulli trials and π_i the success probability of the i^{th} trial (Wang, 1993). For our purposes, π_i represents the probability that an output i was awarded a particular rating, for example 4*.

The Poisson binomial is a special case of the aggregated compound multinomial model used by Brown and Payne (1986). That paper describes a Dirichlet-multinomial ('compound multinomial') model for the unobserved numbers of voters who switched between each of the major parties from one election to another. In their notation, each election featured the same set of political parties. The model estimates the probability, p_{ijk}, that a voter for party i in constituency k becomes a voter for party j.

Our analogy is rather different: there are J parties at the first election, representing the journals in which the articles are published, but only two parties at the next election: '4*' and 'not 4*'.

Voters are articles, and constituencies are academic institutions. We model the probability that an article published in a particular journal is awarded a 4* rating, or not.

Let x_{ij} denote the (known) number of articles published by institution i in journal j. Let y_{ij} denote the (unknown) number of such articles that attained a 4* rating in the REF, with $0 \leq y_{ij} \leq x_{ij}$ for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Let $y_i = \sum_j y_{ij}$ denote the published number of 4* ratings awarded to each institution and let $x_j = \sum_i x_{ij}$ denote the total number of articles submitted from each journal. Then the marginal totals, $\mathbf{Y} = (Y_1, \ldots, Y_I)$, are aggregated compound multinomial (Dirichlet-Poisson-binomial) random variables with expectation

$$\mathbf{E}(\mathbf{Y}) = \mathbf{P}^T \mathbf{x}$$

and covariance

$$\operatorname{cov}(\mathbf{Y}) = \operatorname{diag}(\mathbf{P}^T \mathbf{w}) - \mathbf{P}^T \operatorname{diag}(\mathbf{w}) \mathbf{P},$$

where $\mathbf{P}$ is the J-vector of journal success probabilities[6], $\mathbf{x} = (x_1, \ldots, x_J)$ is a vector of the number of articles in each journal and $\mathbf{w}$ is a J-vector of weights $w_j = x_j(x_j + \alpha_j)/(1 + \alpha_j)$. Brown and Payne (1986) note that 'election data involve more variability than a multinomial would suggest' and add the α vector of J precision/dispersion parameters to account for this.

[6] More generally, $\mathbf{P}$ is a *matrix* of multinomial probabilities

The variance of a Poisson binomial-distributed random variable is

$$\operatorname{Var}(Y_i) = \sum_j (1 - \pi_j)\pi_j = \sum_j (\pi_j - \pi_j^2),$$

which differs from the variance of the aggregate compound multinomial model only by the w_j term. We notice that as α_j grows large, then (dropping the subscripts for the moment)

$$\lim_{\alpha \to \infty} w = \lim_{\alpha \to \infty} x \left(\frac{x}{1 + \alpha} + \frac{1}{\frac{1}{\alpha} + 1} \right) = x,$$

and since $\alpha = \infty$ corresponds to the (non-compound) aggregate multinomial distribution (Brown and Payne, 1986), we can see the Poisson binomial and aggregated multinomial models are equivalent.

If we consider every paper grading to constitute an independent trial, with success probability dependent on the journal in which it is published (but not the institution or any paper-level characteristic), then for each institution i, the number of 4* ratings received is distributed

$$Y_i \sim \text{Poisson-Binomial}(\boldsymbol{\pi}), \tag{6.1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)$ is the vector of journal probabilities[7]. Here the success probabilities are not identical for every publication, but they do coincide wherever two submissions are published in the same journal; if an institution submits two or more articles from the same journal then each of these articles is regarded as a separate independent trial. Of course, independence might be an

[7] In practice, each element π_j is repeated x_{ij} times, representing repeated trials for the number of articles in journal j that were submitted by institution i to the REF.

heroic assumption here, but *ideally* one would hope the REF panels consider each article on its own merits rather than ranking them against one another.

We can fit the model twice: firstly with 'success' defined as 4* ratings, and secondly with success defined as 3* *or* 4* ratings, i.e. 3* or better. As the star ratings are ordinal responses (4* is better than 3*, which is better than 2* and so on), it seems reasonable to assume cumulative odds, and infer the probability of 3* from the estimated probabilities of 4* and of 3*-or-better. Thus a journal's probability of obtaining a 3* rating is assumed to be

$$\pi_j^3 = \pi_j^{34} - \pi_j^4,$$

where we introduce superscript notation: π_j^3, π_j^4 and π_j^{34} respectively denote journal j's probability of accruing 3*, 4* and $\geq 3^*$ ratings.

This separate fitting of the model for the 4* and 3*-or-4* is a key difference from the work of Yan (2017), which used a cumulative odds model with common thresholds for every journal. That is, under Yan's model, $\text{probit}(\pi_j^{34}) - \text{probit}(\pi_j^4) = c$, a constant offset that does not depend on the journal j. Our approach replaces c with c_j, a difference that can might be distinct for every journal.

On top of the likelihood (6.1) we impose a prior on the journal success parameters,

$$\pi_j \sim \text{Beta}\big(\gamma\mu, \gamma(1-\mu)\big) \tag{6.2}$$

for each journal j, such that the mean probability of success is μ, and γ is a regularizing concentration parameter. On top of these we impose hyper-priors

$$\begin{aligned} \mu &\sim \text{Uniform}(0,1) \\ \gamma &\sim \text{Gamma}(\tfrac{1}{10}, \tfrac{1}{20}), \end{aligned} \tag{6.3}$$

where the given hyper-parameters of the gamma distribution are the shape and rate, respectively—corresponding to a mean of 2 and variance 40. In principle one could set these manually, for instance setting μ equal to the empirical mean institutional profile, but we shall try to learn them from the data.

There are more differences between institutions than just the journals in which they publish, so to check for aggregation bias, we extend the model (6.1) such that an article success depends not just on the journal parameter, but on an institutional covariate linked to the REF Environment profiles. In this way we might hope to detect any institutions that perform better or worse in output scores due to the quality of their research environment rather than on the journals in which they publish. Thus the success probability of an article from institution i in journal j is

$$\text{log-odds}(4 * |i,j) = \text{logit}\,\pi_j + \alpha\ \text{envir}_i \tag{6.4}$$

where α is a parameter to be estimated and envir_i is the proportion of 'Environment' in institution i rated 4* (centred by subtracting the

mean). If the α is near zero, then we might conclude that output profiles depend more on the journals than on the unique characteristics of each research institution.

6.4 Methods

We will employ two different methods to estimate the parameters of the model. Firstly, a Bayesian Monte Carlo method, and secondly a maximum likelihood approach using an expectation–maximization algorithm. This section describes the details behind each technique.

6.4.1 Hamiltonian Monte Carlo

Owing to the limited computational power available at the time, Brown and Payne (1986) employed a normal approximation to the Poisson binomial model to estimate the unknown coefficients. The Poisson binomial distribution can also be approximated by a Poisson distribution, though the performance of this approximation is poor when the number of trials is large (Hong, 2013).

Advances in computation capacity allow us to consider a couple of different approaches of fitting a Poisson binomial model. The first would be to employ the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2018) to sample from the posterior distribution via Hamiltonian Monte Carlo (also known as hybrid Monte Carlo or HMC; Duane et al., 1987). Because enumerating all possible sets of integers F_k is computationally infeasible, instead one can program a routine to compute the mass by enumerating with a recursive formula (Shah, 1973)

$$\Pr(K = k) = \begin{cases} \prod_{j=1}^{n}(1 - p_j) & k = 0 \\ \frac{1}{k}\sum_{j=1}^{k}(-1)^{j-1}\Pr(K = k - j)\sum_{l=1}^{n}\left(\frac{p_l}{1-p_l}\right)^j & k > 0, \end{cases}$$

however this may not be numerically stable for large n (Hong, 2013) unless computed on the logarithmic scale.

Using such a dynamic programming algorithm on the logarithmic scale, we fit the model in Stan and report the results in Section 6.6. As a robustness check, we also consider a maximum likelihood approach, described in the next subsection.

6.4.2 Expectation–maximization algorithm

The expectation–maximization (EM) algorithm can provide alternative maximum likelihood point estimates, albeit without any covariance estimate as a measure of uncertainty.

The EM algorithm makes use of the *extended multivariate hypergeometric distribution*. Recall the more familiar hypergeometric distribution describes the probability that, given an urn of N balls, K of them white and $N - K$ black, that if we draw n balls at random then

k of them are white. The *extended* hypergeometric distribution, also known as Fisher's *noncentral* hypergeometric distribution, extends this scenario to non-uniform sampling—where the white balls are more likely to be drawn than black ones because of differences in size or weight.

A multivariate hypergeometric distribution generalizes to a situation where there are more than two colours of balls and describes the probability of picking a particular mixture of colours. Hence, an extended multivariate hypergeometric distribution describes the probability of picking a certain mix from an urn of balls whose weights are not all equal (McCullagh and Nelder, 1989, pp. 260–261). For dimension d different colours, the probability mass function for drawing a mixture $\mathbf{x} = (x_1, \ldots, x_d)$ of n balls is

$$f(\mathbf{x}; n, \mathbf{m}, \boldsymbol{\omega}) = \frac{1}{P_0} \prod_{i=1}^{d} \binom{m_i}{x_i} \omega_i^{x_i}$$

where $\mathbf{m} = (m_1, \ldots, m_d)$ is the number of each colour of balls in the urn and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ are their respective weights. The denominator P_0 is

$$P_0 = \sum_{\mathbf{y} \in \mathcal{S}} \prod_{i=1}^{d} \binom{m_i}{y_i} \omega_i^{y_i}$$

with $\mathcal{S}$ denoting the set of all possible non-negative integer d-vectors $\mathbf{y} = (y_1, \ldots, y_n)$ where $\sum_{i=1}^{d} y_i = n$.

If the article-level ratings were known, we could fit a Rasch-like generalized linear model of the following form to estimate journal effects:

$$\text{logit } \mathbb{E}[\Pr(4^*|i, j)] = \mu + \alpha z_i + \beta_j, \tag{6.5}$$

for a paper by institution i in journal j, where parameter μ, analogously to (6.1), acts as a 'grand mean' intercept term, here in the logit space, and where z_i is an indicator variable for a 'pseudo-institution'. The latter submits an equal number of 4* and not-4* papers from every journal, acting as a regularizing prior on the strength of the journal effects to avoid overfitting. The more pseudo-papers submitted, the stronger the effect of the regularization. The maximum likelihood estimate for such an artificially augmented dataset is equivalent to the posterior mode with a conjugate Bayes model (Jannarone et al., 1990). The optimum strength of regularization (i.e. the number of pseudo-articles augmenting the data) is determined via cross-validation, described at the end of this section.

We adopt an expectation–maximization procedure as follows.

1. Initialize the weights of the noncentral multivariate hypergeometric distribution. That is, randomly generate a probability for each journal–institution that corresponding outputs will be 4*-rated in the REF. (We use a logit-normal distribution for this.)

2. Compute the (approximate) expectation of the noncentral multivariate hypergeometric distribution with these odds (for this,

we use R package `BiasedUrn` by Fog, 2015). This vector forms an imputation of the latent individual-level ratings.

3. Fit the model described in Equation (6.5). Extract coefficients from this model to get new odds for the noncentral multivariate hypergeometric distribution.

4. Repeat steps 2–3 until convergence.

To obtain new odds form the logistic regression model for the noncentral hypergeometric distribution, we simply use the relation

$$\text{odds}(4^*|j) = \exp(\hat{\mu} + \hat{\beta}_j)$$

for all journals j, where $\hat{\mu}$ and $\hat{\beta}_j$ are the estimated parameters from the previous EM step.

Our chosen prior for this model is essentially uninformative on the expected journal ranking. In principle, one could attempt to elicit distributions for the relative strengths of the journals, or (by asking someone who might have served on REF/RAE expert panels in the past) the probability that papers in a certain journal might accrue 4* ratings. However, such an approach is not very scalable to large numbers of journals or fields, so we do not adopt it here.

The cross-validation procedure works as follows.

1. Randomly divide the institutions into (say) 10 groups.
2. For each group:
   a. Run the above expectation–maximization algorithm on data from the other 9 groups.
   b. Use the estimated journal parameters to predict the institutional results for the held-out group.
   c. Compute the index of dissimilarity between the predicted and actual institutional results.
3. Repeat steps 1–2 for different levels of regularization.

We seek the parameter that minimizes the index of dissimilarity (described in the next section) between the predicted institutional scores and the actual scores of the held-out institutions.

6.4.3 Diagnostics and summary statistics

To obtain a 'prediction' or fitted value from the Poisson binomial model, we take the posterior median of the journal probability estimates p_j and take them to be the proportion of the time that articles in those journals were awarded 4*.

That is, we compute

$$\hat{y}_i^4 = \sum_{j=1}^{J} n_{ij} \hat{\pi}_j^4$$

from the model fitted to 4* outputs and

$$\hat{y}_i^{34} = \sum_{j=1}^{J} n_{ij} \hat{\pi}_j^{34}$$

from the same model fitted to a dataset of 3* *or* 4* outputs, where y_i^{34} denotes the number of an institution's outputs rated 3* or better (i.e. 3* *or* 4*), π_j^{34} represents the probability that articles in journal j are awarded 3* or better, and n_{ij} denotes the number of articles from institution i in journal j. Hence we can compute the predicted number of 3* outputs,

$$\hat{y}_i^3 = \hat{y}_i^{34} - \hat{y}_i^4,$$

for each institution $i = 1, \ldots, I$.

Recall that our main aim is to answer the question: to what extent are REF output profiles a function of journal identities? In other words: given the journals in which an institution published its submissions, can we predict that institution's REF score?

To determine the quality of fit of the Poisson binomial model we adopt the index of dissimilarity (Duncan and Duncan, 1955; Kuha and Firth, 2011), which here represents the proportion of an institution's articles predicted a different rating to that observed in the REF. It is computed using the formula

$$\Delta = \frac{1}{2N} \sum_i \left(|y_i^4 - \hat{y}_i^4| + |y_i^3 - \hat{y}_i^3| + |y_i^4 + y_i^3 - \hat{y}_i^4 - \hat{y}_i^3| \right),$$

where $N = \sum_i \sum_j n_{ij}$, the total number of submitted outputs.

From the index of dissimilarity we propose another metric, the *redistribution of monetary reward*, based on the notion that a 4* output is worth four times as much in research funding as a 3* output, and outputs rated 2* or lower accrue no direct funding at all (see e.g. Koya and Chowdhury, 2017). This metric describes the fraction of total monetary reward that would move between institutions if the estimated REF profiles $(\hat{y}_i^4, \hat{y}_i^3)$ were used instead of the observed profiles (y_i^4, y_i^3), and is measured by

$$\Delta_{£} = \frac{\frac{1}{2} \sum_i m_i |r_4 (p_i^4 - \hat{p}_i^4) + r_3 (p_i^3 - \hat{p}_i^3)|}{\sum_i m_i (r_4 p_i^4 + r_3 p_i^3)},$$

where m_i is the number of full-time equivalent (FTE) staff submitted by institution i in the unit of assessment, $\hat{p}_i^4 = \hat{y}_i^4 / \sum_j n_{ij}$, $\hat{p}_i^3 = \hat{y}_i^3 / \sum_j n_{ij}$ and r_4 and r_3 are the respective monetary reward per FTE for the 4* and 3* components of output profiles, in arbitrary units with $r_4 = 4r_3$. (Implicitly, terms for 2*, 1* and unclassified outputs can appear in the above formula, but we take $r_2 = r_1 = r_u = 0$.)

The monetary index might even be combined with calculations of the kind by Koya and Chowdhury (2017) to compute an absolute sterling figure for the amount of funding that would move institutions in a switch from the actual REF profiles to those estimated our model.

6.5 Data

6.5.1 Units of assessment

To demonstrate the method, we will first consider the 'Economics and Econometrics' unit of assessment. We choose this particular subject area because it is small, relatively self-contained, and a high proportion of output submissions (92%) are in the form of journal articles (rather than books, conference proceedings or other works). One might expect (this being a statistics PhD thesis) to look at statistical science submissions first, however these fall under the umbrella of Mathematical Sciences—along with research in probability, pure and applied mathematics and mathematical physics—which is a larger and more hetereogeneous field.

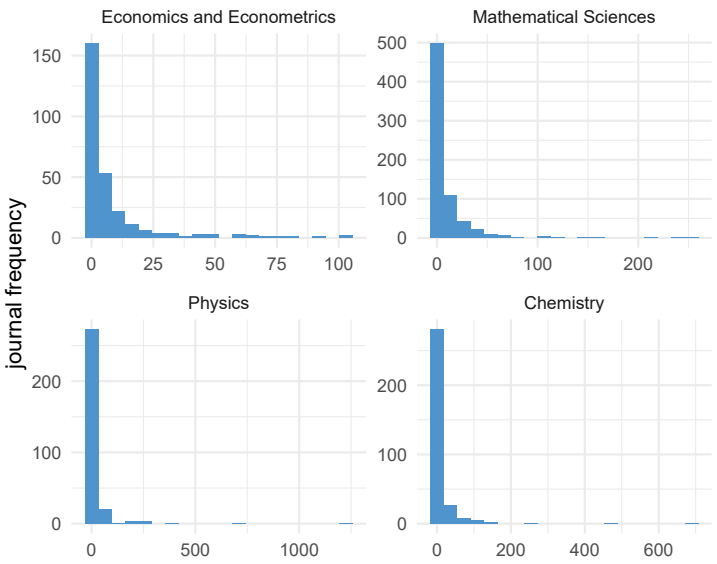
From Table 6.2 it is easy to see that the hard sciences (REF panels A and B) mostly submitted outputs in the form of journal articles; the arts and humanities (panel D) used other formats, and social sciences (panel C) were somewhere in between. A notable exception to this rule is the field of Computer Science and Informatics, where the role of academic journals is often supplanted by conference proceedings.

After the initial analysis of the Economics and Econometrics sub-panel, we also examine three other REF subpanels, all from the Physical Sciences main panel, to see how they compare. The arts (main panel D) publish too few of their outputs as journal articles for this model to be of practical relevance.

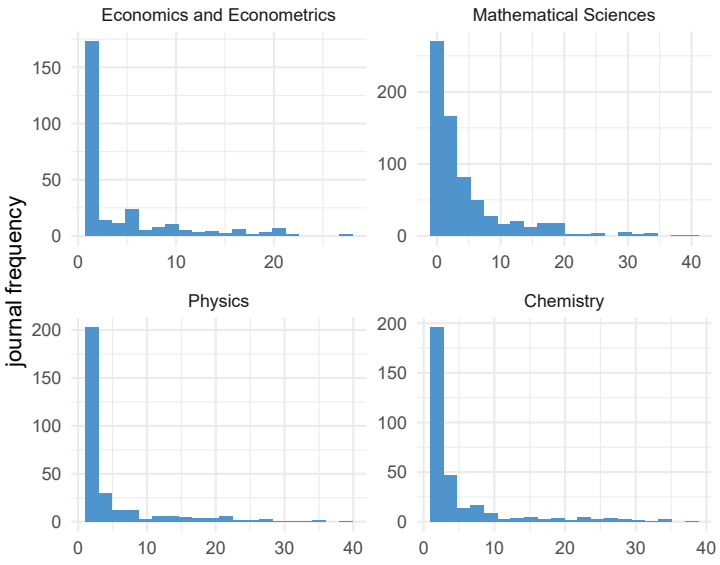
6.5.2 Wrangling REF2014 data

In the published REF2014 submissions data, outputs are explicitly categorised into types, such as journal articles, book chapters, conference proceedings, working papers and so on. However, there is no sure-fire way to group together articles published in the same journal or book, as the titles are unstandardised, ISSNs, if provided, can vary between print and online editions and DOIs, where present, can be difficult to parse. Labour-intensive manual tagging of the data has rather little appeal, not least because it is error-prone and does not scale well to larger future data sets. But there is a network science solution to the problem.

We coerced the output submissions data into a long-format table comprising just a journal identifier—the ISSN, ISBN, DOI or standardised journal title (coerced to lower case, with punctuation, diacritics, spaces and leading "the"s removed)—and a unique identifier for each output, then constructed an undirected bipartite graph between the journal identifiers and individual output identifiers. Each connected component in this graph represents a unique journal, containing outputs with a common journal title, DOI, ISSN and/or ISBN. Each is assigned a unique journal ID, as well as a human-readable title, the latter sampled from one of the journal title variants found in the component.



(a) number of submitted articles



(b) number of submitting institutions

Figure 6.1: Distribution of journal articles across journals and institutions, by unit of assessment. Some journals are much more popular than others, and not all institutions publish in the same journals

Panel	Unit of assessment	Outputs	Journals
A	Clinical Medicine	13400	99.9
	Biological Sciences	8608	99.7
	Public Health, Health Services and Primary Care	4881	99.6
	Psychology, Psychiatry and Neuroscience	9126	99.6
	Agriculture, Veterinary and Food Science	3919	99.1
	Allied Health Professions, Dentistry, Nursing and Pharmacy	10358	98.9
B	Chemistry	4698	99.8
	Earth Systems and Environmental Sciences	5249	99.1
	Aeronautical, Mechanical, Chemical and Manufacturing Engineering	4143	99.0
	Electrical and Electronic Engineering, Metallurgy and Materials	4025	98.9
	Physics	6446	98.9
	General Engineering	8679	98.4
	Civil and Construction Engineering	1384	97.4
	Mathematical Sciences	6994	96.2
	Computer Science and Informatics	7651	72.6
C	Sport and Exercise Sciences, Leisure and Tourism	2757	96.8
	Business and Management Studies	12202	95.6
	Economics and Econometrics	2600	91.8
	Geography, Environmental Studies and Archaeology	6017	82.6
	Education	5519	78.3
	Architecture, Built Environment and Planning	3781	77.6
	Social Work and Social Policy	4784	77.4
	Sociology	2630	76.1
	Politics and International Studies	4365	70.6
	Anthropology and Development Studies	2013	67.3
	Law	5522	62.5
D	Philosophy	2173	61.8
	Area Studies	1724	56.6
	Communication, Cultural and Media Studies, Library and Information Management	3517	52.5
	Modern Languages and Linguistics	4932	48.3
	History	6431	44.0
	Theology and Religious Studies	1558	37.2
	English Language and Literature	6923	35.7
	Music, Drama, Dance and Performing Arts	4246	29.8
	Classics	1386	28.9
	Art and Design: History, Practice and Theory	6321	26.2

Table 6.2: Units of assessment in REF2014, the number of outputs submitted and the percentage of which that were classified as journal articles

Unfortunately, this methodology on the published REF data alone assumes integrity of the published data, which was later found to be lacking. Some administrators entered article metadata by hand, rather than retrieving it programmatically via CrossRef, as perhaps they should have done. This inevitably introduced human error; for example one entry that should have been from the *Annals of Mathematics* had the correct DOI, article and journal title, but the ISSN was that for the separate *Advances in Mathematics* journal, which causes the above mini-algorithm to merge the works in *Annals* and *Advances* as if they came from the same journal. In turn, the *Advances in Mathematics* journal was grouped with *Advances in Applied Mathematics* due to similarly careless data entry. Further issues were caused by journal titles that were ambiguous if not completely erroneous, for example various articles published in

Physical Review Letters, *Physical Review A*, *Physical Review B* and so on all being given the unhelpful abbreviation *PHYS REV*.

Evidently, the metadata in the published REF outputs data set cannot be trusted, except possibly the DOIs. To remedy this, we used the R package **rcrossref** (Chamberlain et al., 2019) to access CrossRef application programming interface (API), allowing retrieval of metadata associated with the 25,000 unique DOIs for the Economics & Econometrics, Mathematical Sciences, Chemistry and Physics submissions. All except 22 returned results. Of these few 'invalid' DOIs, manual inspection showed the same broken DOIs to be published on publishers' own web pages (and this was reported to CrossRef) so these were not a problem with the REF data itself. For the remaining (vast majority) of DOIs, the CrossREF API returned the titles of the articles and the names and ISSNs of the containing periodicals.

A small amount of data wrangling remained, however. Though no single DOI yielded multiple entries in the CrossREF database, our mini clustering algorithm was still required to merge journals which have multiple titles appearing in CrossRef, for example *The Review of Economic Studies* and *Review of Economic Studies*. These were able to be clustered by shared ISSNs (and we assume that CrossRef, at least, gets these correct).

This approach can easily be applied to every field with no manual or *ad hoc* data processing necessary (except those articles with missing or invalid DOIs). The distribution of outputs to journals and to institutions is illustrated in Figure 6.1, where we can see that it is quite skewed. An uneven spread of journals between institutions is desirable for an ecological inference model; if every institution published in the same profile of journals then it would be impossible to learn any journal-level effect.

Nevertheless, estimating several hundred journal parameters from just a few dozen institution-level observations is particularly ambitious, especially when it is evident that many journals accounted for very few submitted outputs.

Ordinarily in high-dimensional data analysis, one can apply some level of regularization to the model, the exact level of regularization to be determined by, say, empirical Bayes estimation. However, standard techniques of 'soft' regularization do not seem to work very well for aggregated data like those found in our ecological inference problem. Instead we adopt a fairly pragmatic approach: any journal containing fewer than some threshold number of articles is aggregated into a single *super-journal* entitled 'Other journals'. We choose the threshold such that *most* (i.e. $\geq 50\%$) of the articles in the data fall into a named journal rather than an anonymous 'other' journal, while hopefully also keeping the number of parameters low enough to be practical for reporting and visualization. Conference proceedings and other non-journal outputs cannot be ignored, as the Poisson binomial model requires we account for all submitted outputs, so these publications are aggregated into

their own respective 'other' categories.

We apply our methodology to the Economics and Econometrics sub-panel as well as three other fields: Mathematical Sciences, Physics and Chemistry, representing three units of assessment from REF2014 main panel B. Biological Sciences (main panel A) was also considered, but modelling this field proved too computationally intensive, possibly due to the large number of submitted outputs (8,608) and institutions (44) or the distributions thereof. (This unit of assessment could still be analysed in future with a more efficient model fitting implementation.)

Compared to Economics and Econometrics, several times more outputs were submitted to each of these sub-panels (see Table 6.2), the vast majority of them ($\geq$ 96%) in the form of journal articles.

6.5.3 Economics & Econometrics

Our first REF sub-panel of interest, the 'Economics and Econometrics' unit of assessment, received 2600 publications from 28 institutions for the outputs submission. Of these, 2388 were journal articles, distributed in various publications as shown in Table A.1.

Using a combination of CrossRef data and the clustering algorithm described in the previous section, eventually we were able automatically to assign the 2388 economics outputs into 277 unique journals.

Setting the threshold at all Economics and Econometrics journals containing $\geq$ 20 submitted articles, we obtain the distribution shown in Table A.1. There are 29 named journals, representing over half of the total submissions.

6.5.4 Mathematical Sciences

After the field of economics, we study the Mathematical Sciences unit of assessment, which encompasses pure and applied mathematics, statistics and probability—though no distinction was made between these sub-fields, so the REF panel perhaps had the dubious honour of trying to assess subfields as diverse as pure mathematics and applied statistics together on the same measurement scale.

Mathematical Sciences was larger than Economics & Econometrics, with 53 submitting research institutions. The 6994 Mathematical Sciences outputs, of which 6731 were classified as journal articles, span some 696 unique scholarly journals. In this case, and for the remaining three sub-panels, the larger number of articles per journal necessitates a higher threshold for named journals: we increase the minimum number of submitted articles to 30 for Mathematical Sciences, Physics and Chemistry. This ensures that 'named' journals still provide a good representation ($\geq$ 50% coverage) of outputs in the data, while keeping model complexity reasonably low.

Figure 6.1a suggests a similarly skewed distribution of articles across journals: many journals represented just one or two article

submissions each, but a small number of mainly physical science journals had article counts in triple figures, including *Journal of Fluid Mechanics* with 254 articles and *Physical Review Letters* with 209. Some sub-fields appear to have published (or at least been submitted) more prolificly than others: the biggest statistical journal submission number was from *Biometrika* with 57 articles. See Table A.3 for a full breakdown.

Across institutions, the journal submissions data for Mathematical Sciences are skewed: most journals were published in by only a handful of unique institutions, but there were a small number of journals popular with nearly all of the institutions assessed by the sub-panel. See Figure 6.1b.

6.5.5 Physics

We now turn to Physics, with 6446 REF2014 outputs, of which 6376 were journal articles in 304 unique journals, which we might expect to have some overlap with the Mathematical Sciences. Indeed, some 100 journal titles appear in both submissions.

There were 41 different institutions who submitted to the Physics sub-panel for REF2014.

As with Mathematical Sciences, in Physics we used a cut-off of 30 articles for a publication to be 'named' in the model, rather than aggregated under 'Other journals'.

The distribution of journals by article count and across institutions, shown in Figure 6.1, appear largely similar to the aforementioned subjects, but the breakdown of article counts by journal in Table A.2 reveals that two journals—*Physical Review Letters* and *Monthly Notices of the Royal Astronomical Society*—were extremely strongly represented, constituting nearly 30% of all outputs.

6.5.6 Chemistry

Our fourth unit of assessment to model is Chemistry. The data for this field comprise 37 institutions, who submitted 4698 outputs, of which 4688 were journal articles in 326 unique journals.

The distributions of submissions, shown in Figure 6.1, once again look similar to the other fields. As in Physics, a couple of journals stand out for containing a very high proportion of outputs: the *Journal of the American Chemical Society* and *Angewandte Chemie* together represent nearly 25% of all submitted works.

6.6 Results

Figure A.1 represents the posterior marginal density for the parameter α, defined in (6.4) as the effect of institutions' research environments—rather than journal submissions—on the probability of their outputs attaining 4* ratings in the REF.

For Economics and Econometrics, Figure A.1a suggests there is little evidence for the environmental effect being distinct from zero,

either when estimating the probabilities of journals attaining 4* or $\geq 3^*$ ratings in the REF. The same was found also for Mathematical Sciences, Physics and Chemistry. This simple diagnostic check—for sensitivity of the results when controlling for an institution-level covariate—provides some, albeit limited, reassurance: the results appear robust to potential effects of aggregation bias, and there is no indication from this check of anything like Simpson's paradox.

Trace plots for the Hamiltonian Monte Carlo runs are given in Figure A.2, and suggest good mixing of the chains for each of the parameters.

To catch any glaring errors in the results, and for a more informed interpretation of the findings (especially the implied journal rankings in each field) the initial results were presented to several senior University of Warwick academics with expertise in their respective disciplines. This was invaluable, for example, in spotting the conspicuous absence of *Annals of Mathematics* from the rankings, due to the aforementioned coding error in the REF2014 data. With such anomalies fixed, our informal panel of experts provided useful context for the final results, presented in the following sections.

6.6.1 Economics & Econometrics

Figure 6.2 provides a 'league table', in the form of a series of box plots of the marginal posterior distributions, of the estimated Economics and Econometrics journal probabilities of attaining 4* and 3* or 4* ratings in the REF. Strikingly, the five journals considered among economists to be the 'Top Five' in their field (Heckman and Moktan, 2018) are near the top of this ranking as well: namely, the *American Economic Review*, *Econometrica*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and even the *Journal of the Political Economy*, despite the latter only representing a handful of outputs, at 22 articles. Looking at the probability of achieving 3* or greater (Figure 6.2b), the top probabilities are all so close to 1 that little can be inferred from the ordering of the journals.

The 95% posterior intervals are quite wide, especially for publications with fewer articles submitted in the REF, which is to be expected given the inherent uncertainty associated with estimating a large number of parameters from a small number of incomplete observations.

Our journal ranking has several notable omissions: the *Journal of Labor Economics* and the *RAND Journal of Economics* are highly respected (Sgroi, 2019; Oswald, 2019), as are general science journals such as *Science*, *Nature* and *PNAS*. However none of these journals met the minimum threshold of 20 articles submitted to the REF2014 Economics & Econometrics sub-panel, so they do not appear as 'named journals' in our results.

As a robustness check, Figure 6.3 compares Hamiltonian Monte Carlo estimated journal probabilities of attaining 4* with the respective maximum likelihood estimates computed via the expectation–

maximization algorithm (with the level of pseudo-data set (arbitrarily) at one article per journal). The maximum likelihood estimates come without any uncertainty quantification, but we can see a strong correlation in point estimates between the β_j estimates of Equation (6.5) and the (logit) success probabilities corresponding to the same journals, so the general approach seems sound.

Figure 6.4a shows the predicted versus actual 4* output profiles for each of the institutions in Economics and Econometrics. With the predicted 4* and 3* profiles converted into funding allocations, Figure 6.4b shows the resulting discrepancies between the predicted institutional funding versus that actually allocated (based on the methodology of Koya and Chowdhury, 2017) based on HEFCE data. Not all institutions are based in England, of course, so the 'actual' funding figures for other nations in the UK assume that the respective research councils used similar formulae to allocate funding based on REF2014 outputs.

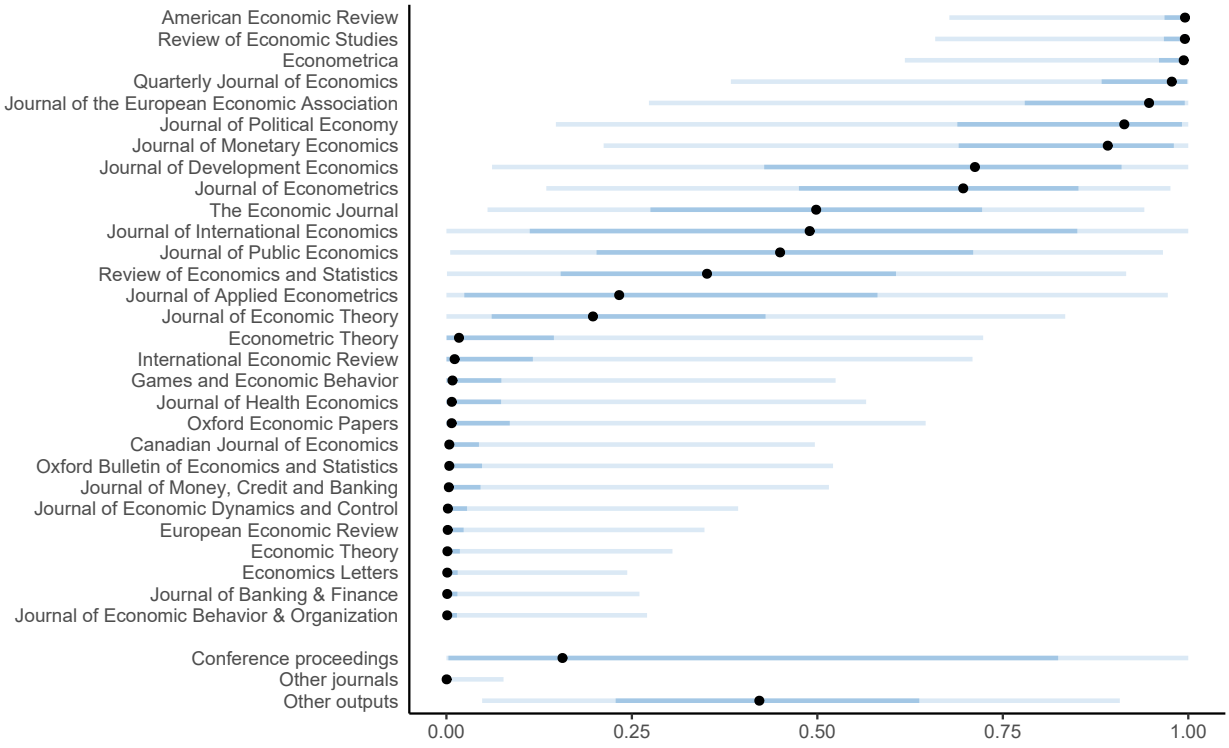
The quality of prediction appears reasonably good, with most points falling close to the line of $y = x$. Some institutions appear to have received more 4* ratings than predicted from their journal choices, notably Cambridge and UCL, and Queen Mary University London appears to have received fewer 4* ratings than suggested by the model. Otherwise there are no noticeable outliers.

When it comes to funding, Figure 6.4b shows how much funding would be allocated, when combining the estimated 4* profiles with 3* ratings and the FTE headcount for each department. The only institution with a significant discrepancy is Brunel, and this can be accounted for by the fact that most of that university's Economics and Econometrics outputs were published in less popular journals not named in Table A.1, so the model has less information available to predict this institution's results.

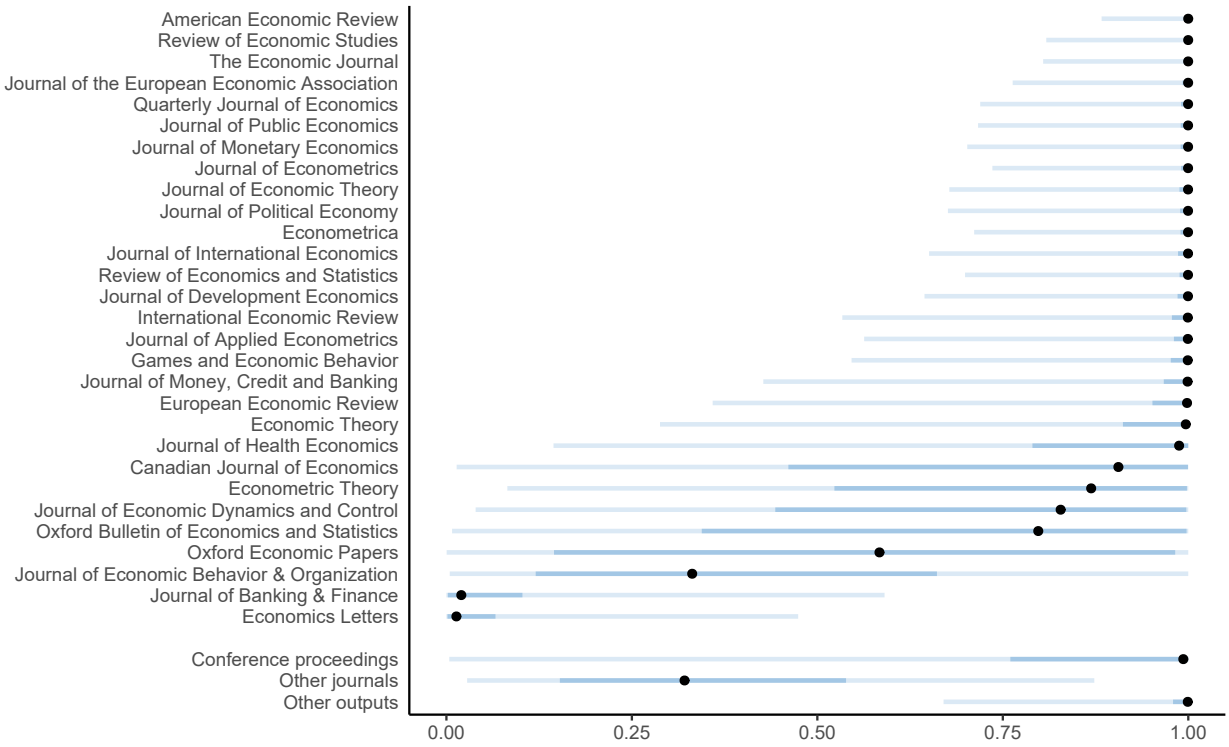
Across the four fields, we compute the index of dissimilarity, Δ, and the index of redistribution of monetary reward, $\Delta_{£}$. The distributions of these metrics are plotted in Figure 6.5. Lower numbers are better.

The median value for Economics and Econometrics is $\Delta = 17.9\%$, that is, this proportion of articles would need to be reclassified for the estimated institutional profiles to match exactly those published in the REF. As a metric, 82.1% accuracy sounds like it might be quite good, but we should be careful not to draw too many conclusions from a single number. In funding terms, that translates to $\Delta_{£} = 8.7\%$ of funding in Economics and Econometrics needing to be reallocated if an initial allocation was made based on the Poisson binomial model alone. Across dozens of institutions, that represents a substantial amount of money, though.

Figure 6.6 provides evidence against the model of Yan (2017), which assumed a constant cumulative probit difference between the probability of getting 4* and the probability of getting 3* or better. It is clear that 'better' journals (those more likely to attain 4*) have a smaller cumulative probit difference, suggesting that it is not



(a) Probability of 4*



(b) Probability of 3* or 4*

Figure 6.2: Median estimated journal success probabilities in Economics and Econometrics. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 20 or more articles submitted in REF2014

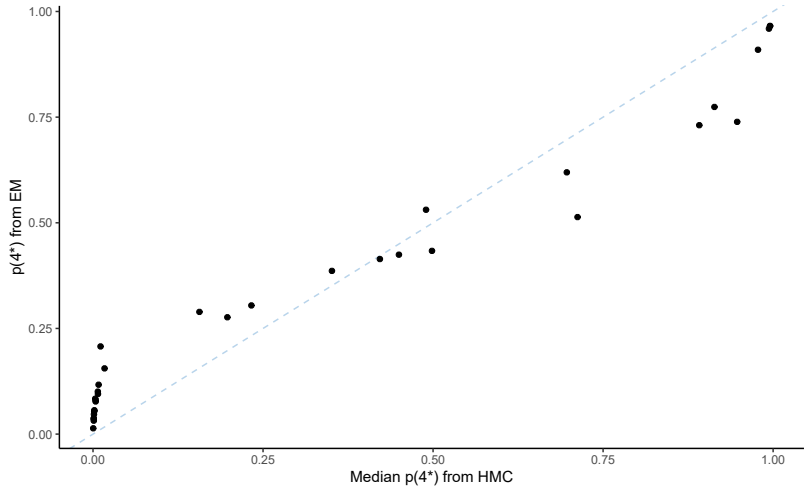


Figure 6.3: Maximum likelihood estimates of journal effects, $\hat{\beta}_j$, versus Hamiltonian Monte Carlo estimates of journal success probabilities (on a logit scale), for Economics and Econometrics, with line of best fit

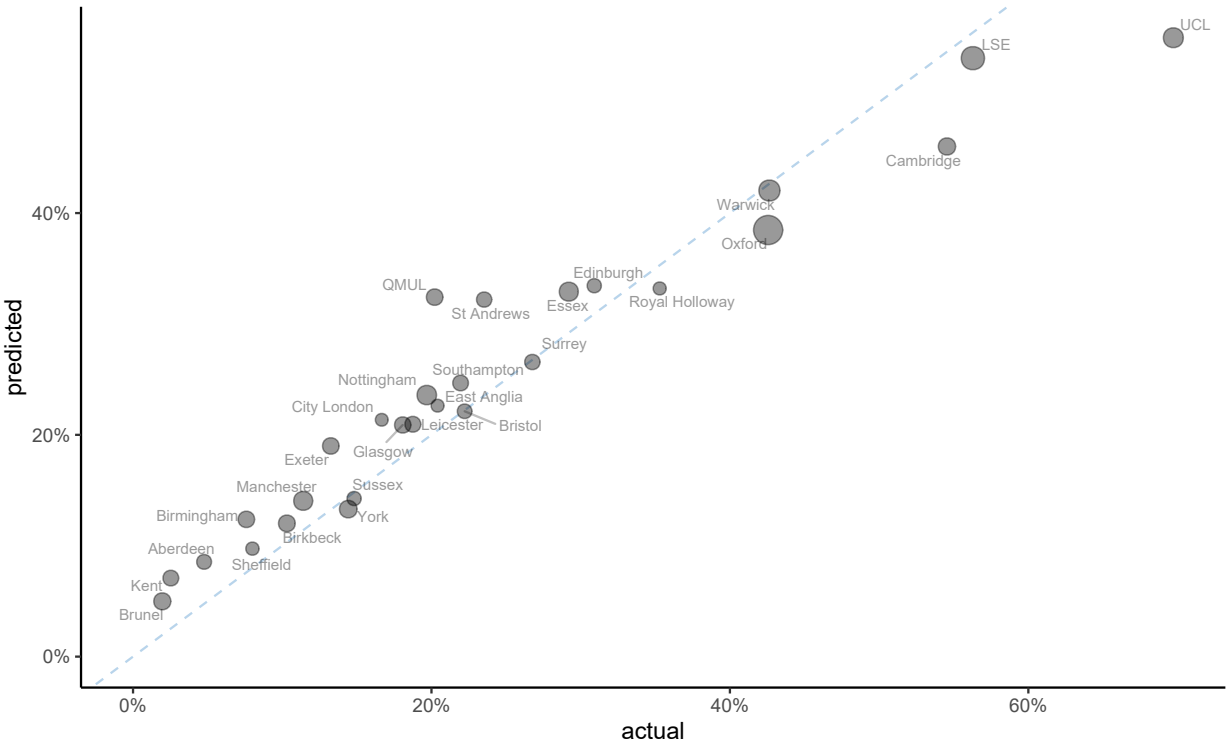
much harder for an output in such an apparently high-achieving journal to get a 4* than a 3* rating, whereas for 'weaker' journals, it is harder to improve from 3* to 4*.

6.6.2 *Mathematical Sciences*

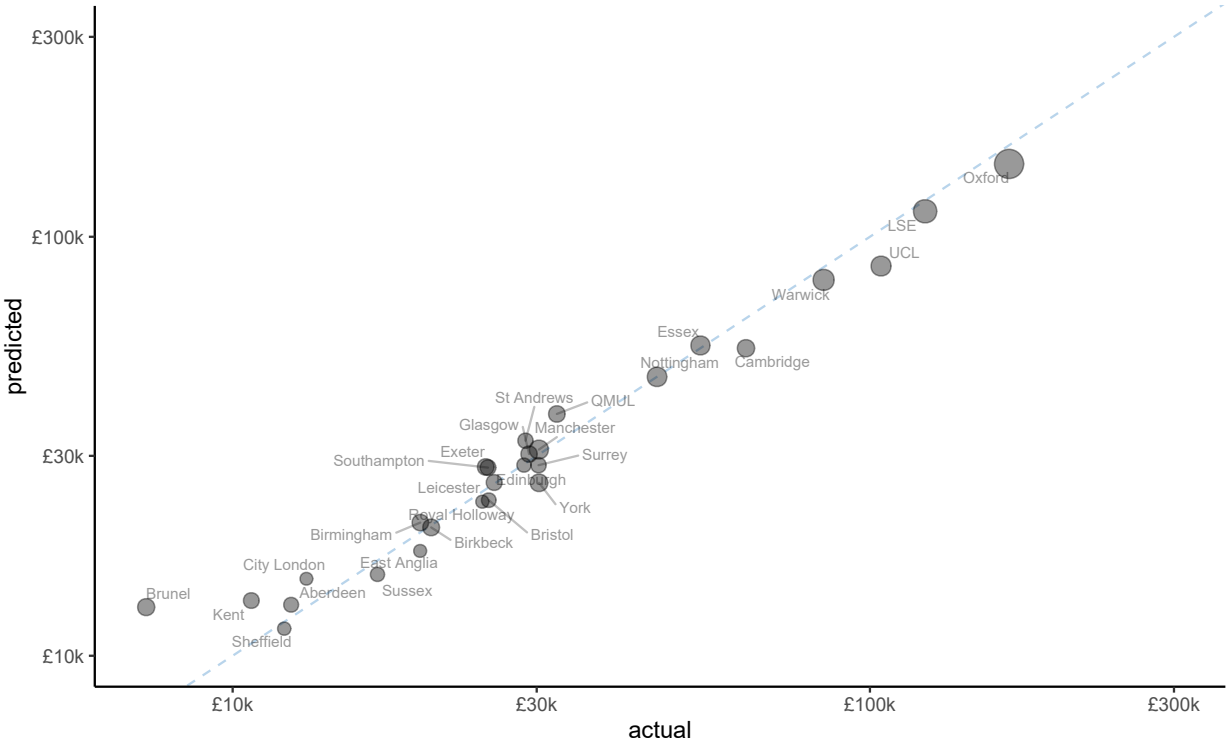
In Mathematical Sciences, we face the problem of several partly disjointed sub-fields, such as pure mathematics, statistics, mathematical physics and mathematical biology, all falling under the same umbrella. As a result it is harder to gauge what might be considered a group of 'top' mathematical sciences journals—mathematicians might declare that statistics, for example, is merely applied mathematics and that a pure mathematics journal should lead the field (Monroe, 2008) whilst statisticians might counter that statistical journals should come top because of the widespread application of statistics. It is perhaps surprising, then, that some of the reputed top journals in statistics, *Annals of Statistics*, *Biometrika* and the *Journal of the Royal Statistical Society: Series B* (Varin et al., 2016) are still ranked highly based on the model for attaining 4* ratings in the REF. See Figure 6.7. However, the *Journal of the American Statistical Association*, also a highly-regarded statistics journal, has a low estimated probability of obtaining 4* ratings.

For the mathematicians, *Inventiones Mathematicae* and *Annals of Mathematics* are both highly reputed and have the highest estimated probabilities of yielding 4* ratings in the REF. The *Journal of the American Mathematical Society* and *Publications Mathémathiques de l'IHÉS* are also highly regarded (Loeffler, 2019), but do not appear in the results as named journals because fewer than 30 of their respective articles were submitted to the REF.

In Mathematical Sciences, the predicted versus allocated 4* ratings and funding allocations, by institution, are presented in Figure 6.9. Apparent outliers (such as Coventry University or the University of Greenwich in the 4* plot) are among the smallest institutions by number of full-time equivalent (FTE) research staff.

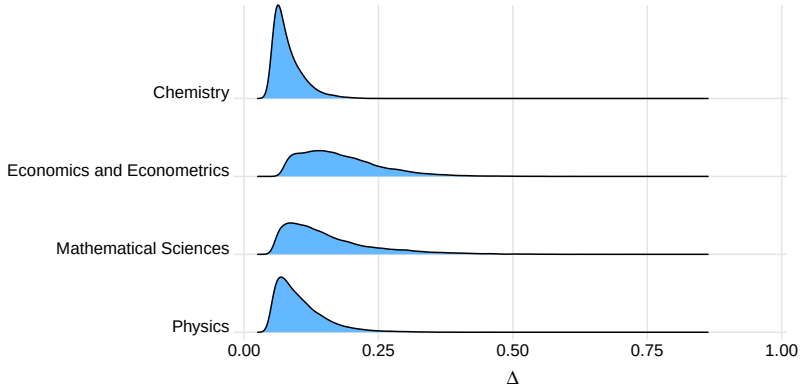


(a) % articles at 4*

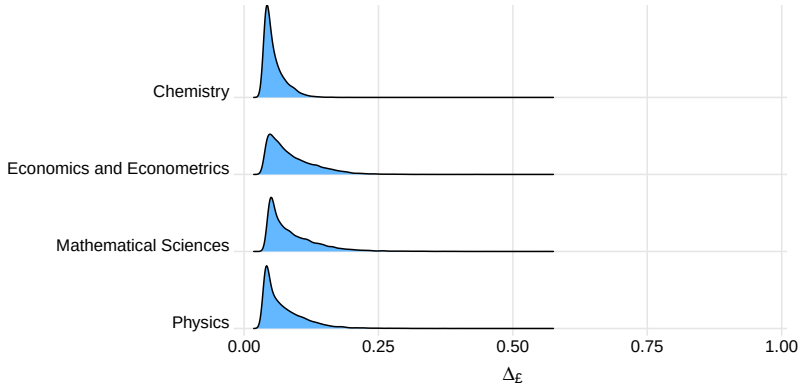


(b) funding allocation

Figure 6.4: Predictions versus observed REF2014 results for institutions submitting outputs to the Economics & Econometrics sub-panel, with point sizes proportional to number of FTE staff



(a) Index of dissimilarity



(b) Index of redistribution of monetary reward

Figure 6.5: Density plots of indices of dissimilarity and of redistribution of monetary reward, by unit of assessment

There appears to be a pattern, however: weaker institutions are expected to do better, and stronger institutions are expected to do worse, than their actual published performance in the REF.

This shrinkage effect implies that some variation in assessed quality of outputs is not explained by journal identities alone. It indicates that there is variation in quality within at least some journals, and that high-ranked institutions tend to publish more of the high-quality papers in such journals.

In terms of summary measures, the median index of dissimilarity for Mathematical Sciences is 15.5% and the median required redistribution of monetary reward is 8.9%; the posterior distributions of these statistics are plotted in Figure 6.5.

6.6.3 *Physics*

Posterior probabilities for the Physics sub-panel are presented in Figure 6.10. Journals from Nature Publishing Group have the highest estimated probabilities of attaining 4*, though no probabilities are near 100%, perhaps owing to the relatively small number of 4* ratings awarded in this field generally. The appearance of *Physics Review Letters* above the *Proceedings of the National Academy of Sciences (PNAS)* in the ranking might imply a preference by physicists in the review panel for physics-specific journals over general science ones. In the international astrophysics community, *Astro-*

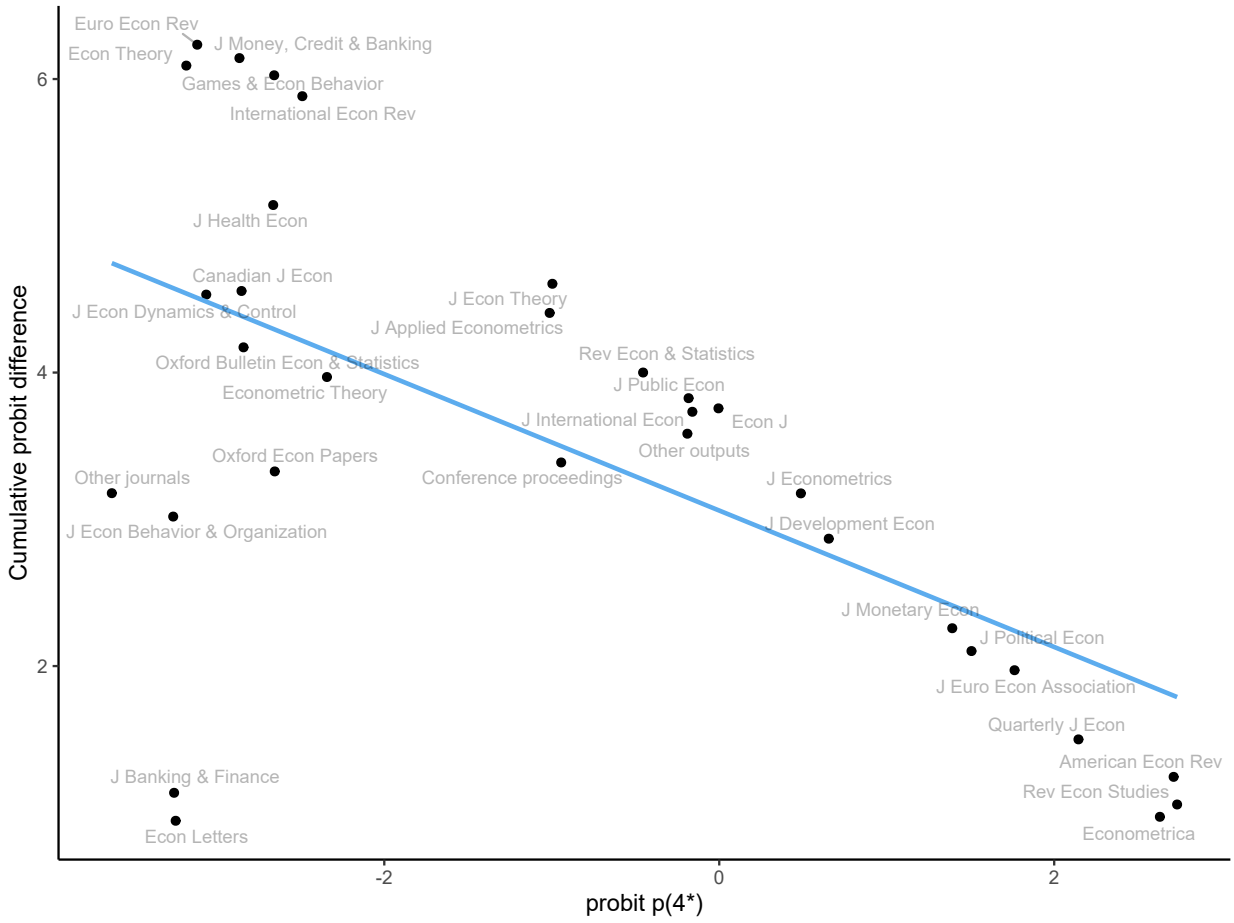


Figure 6.6: Comparison of cumulative probit differences, $c_j = \text{probit}(p_j^{34}) - \text{probit}(p_j^4)$, versus estimated probit probability of attaining 4*, by journal in Economics and Econometrics in REF2014, with line of best fit. A non-zero slope implies $c_j \neq c$, that the cumulative probit difference is not constant across journals

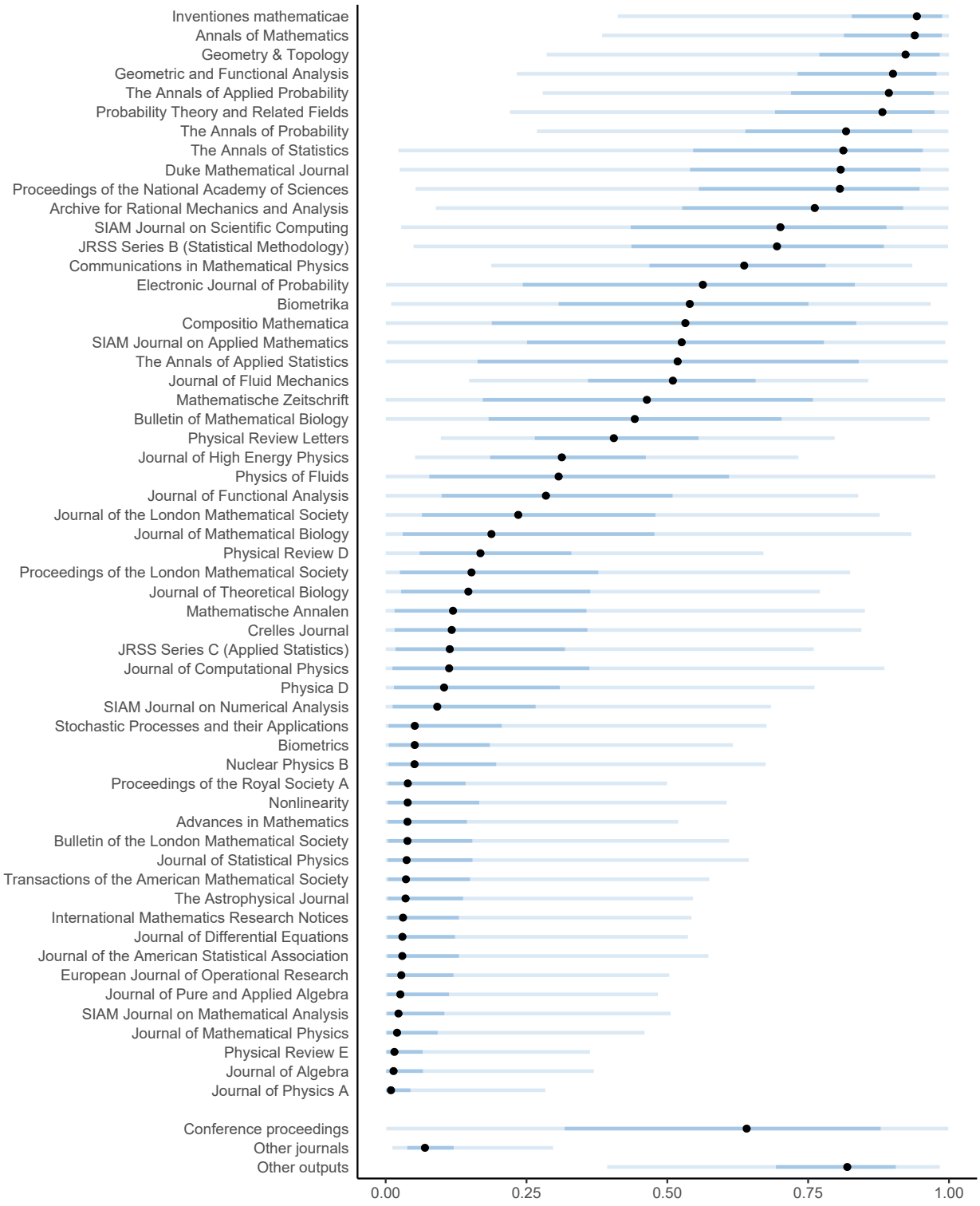


Figure 6.7: Median estimated journal success probabilities of 4* ratings in Mathematical Sciences. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014

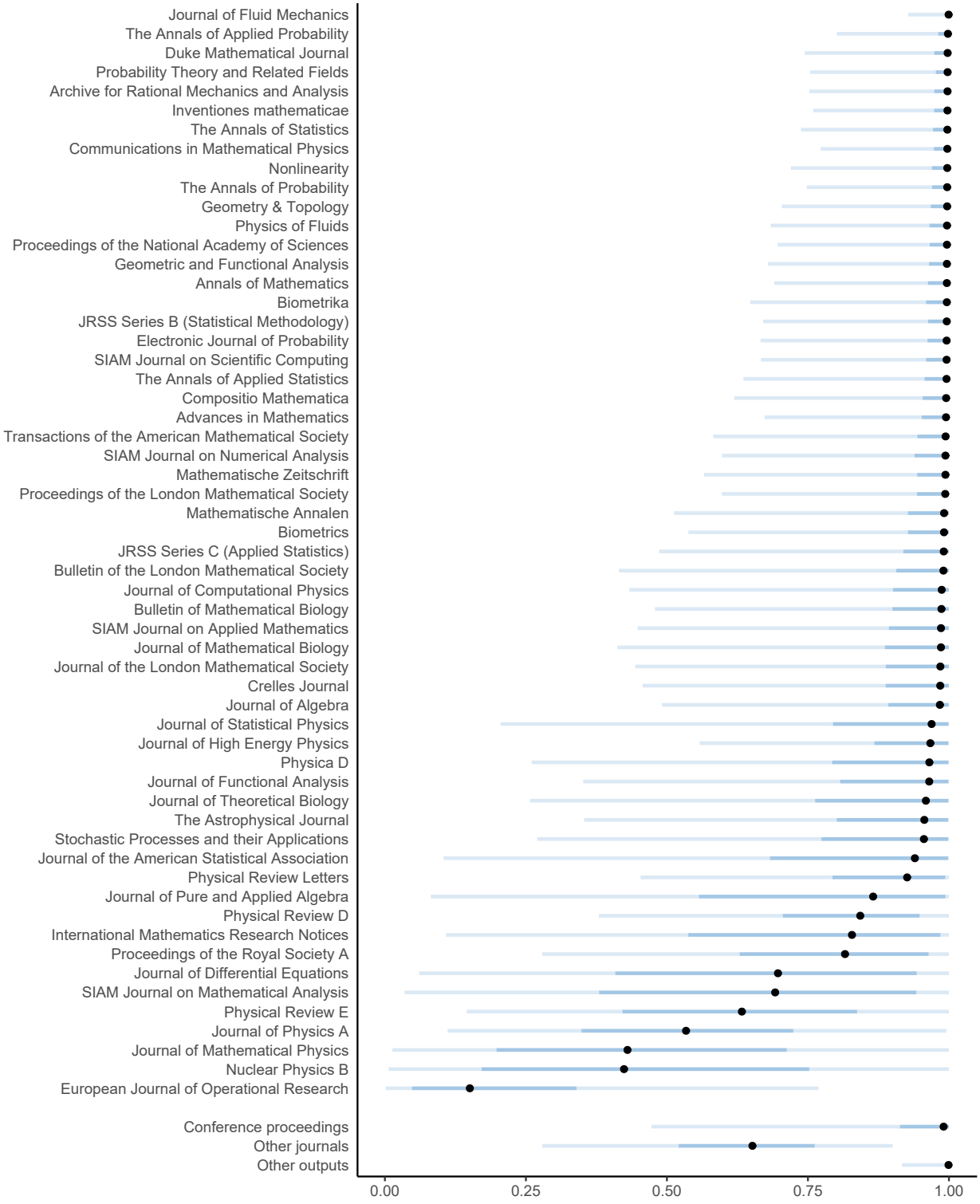
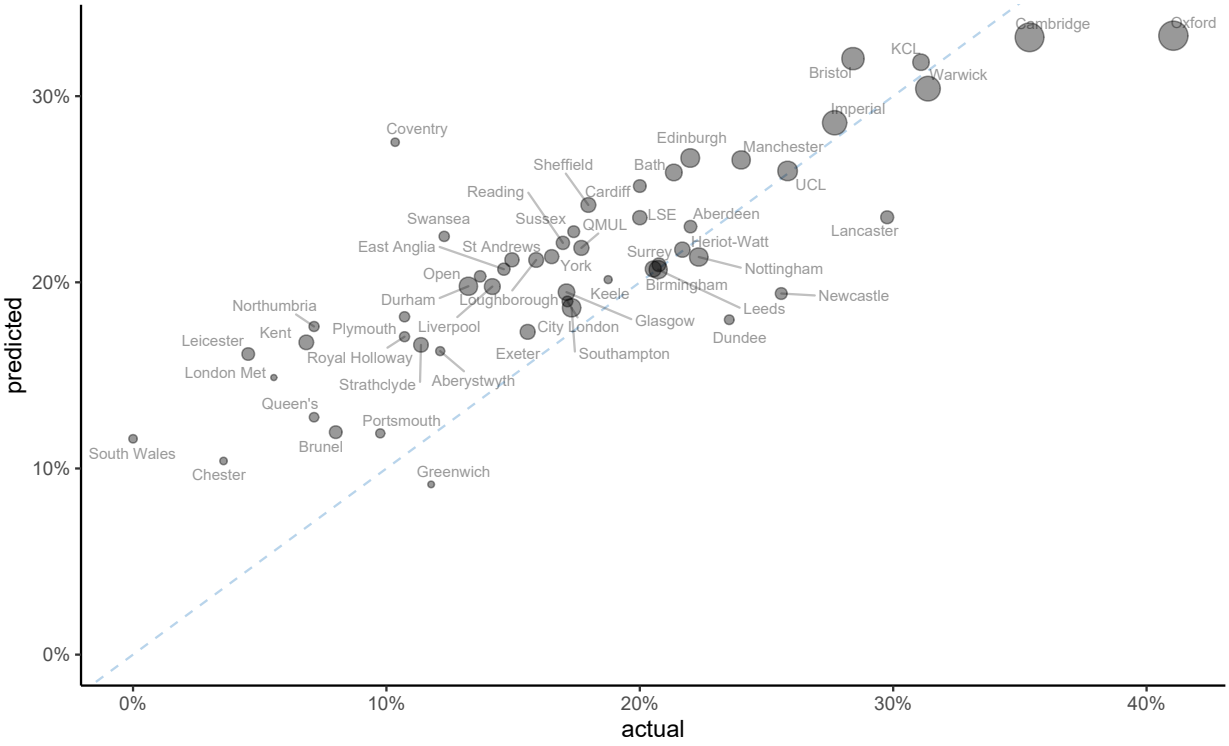
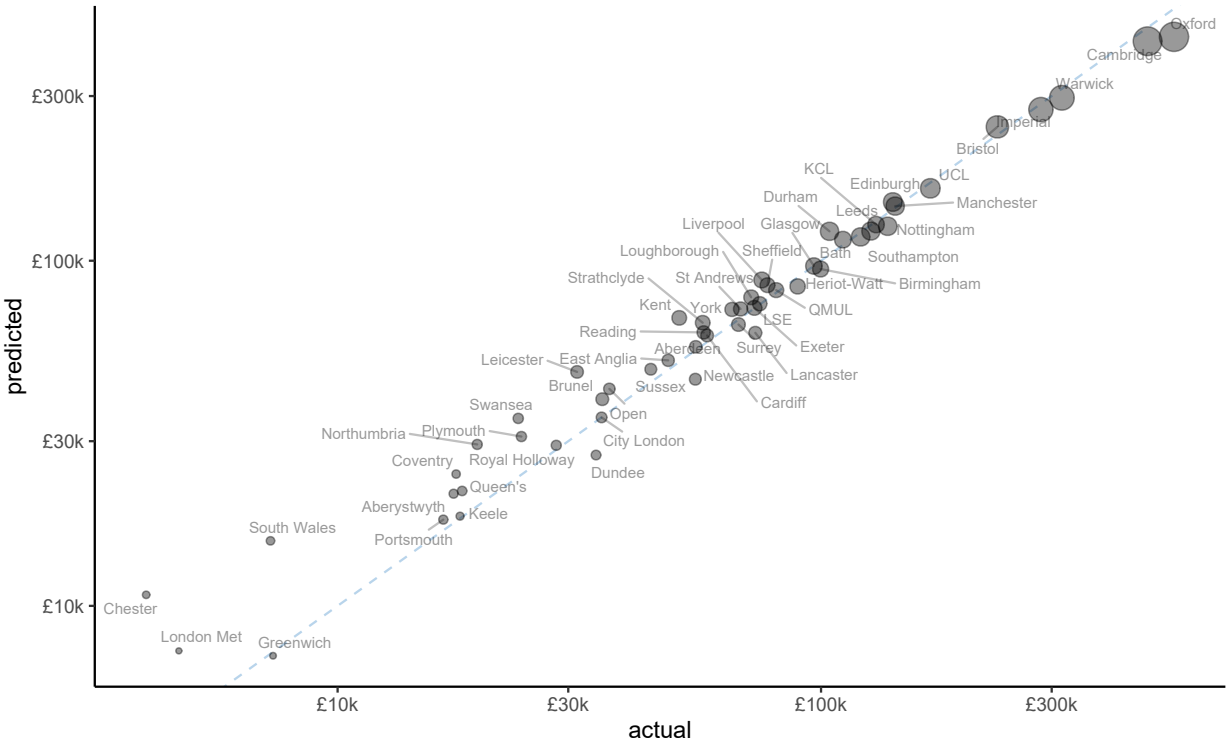


Figure 6.8: Median estimated journal success probabilities of 3* or 4* ratings in Mathematical Sciences. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014



(a) % articles at 4*



(b) funding allocation

Figure 6.9: Predictions versus observed REF2014 results for institutions submitting outputs to the Mathematical Sciences sub-panel, with point sizes proportional to number of FTE staff

physical Journal might be considered more prestigious than *Monthly Notices of the Royal Astronomical Society*, but the latter has a slightly higher estimated probability of 4*, which might be interpreted as a UK-centric bias (Ball, 2019). Relatively low success probabilities for *Physics Review B* and *C* could be attributed to an inter-journal dependence: namely, some works published in these journals also being announced in the highly-ranked *Physical Review Letters*.

As in Mathematical Sciences, a comparison of the predicted versus actual institutional REF results in Physics, shown in Figure 6.11, reveals a linear relationship, but an apparent shrinkage effect, implying some variation in assessed quality not explained by journal identities. The performance of the University of Oxford, in particular, appears to be under-estimated by the model, suggesting that where there is variation of assessed quality within journals, the higher-quality outputs may be more likely to have been from Oxford researchers.

By summary measures, the median index of similarity in Physics is 10.5% and the median proportion of reallocated research funding is 7.3%.

6.6.4 Chemistry

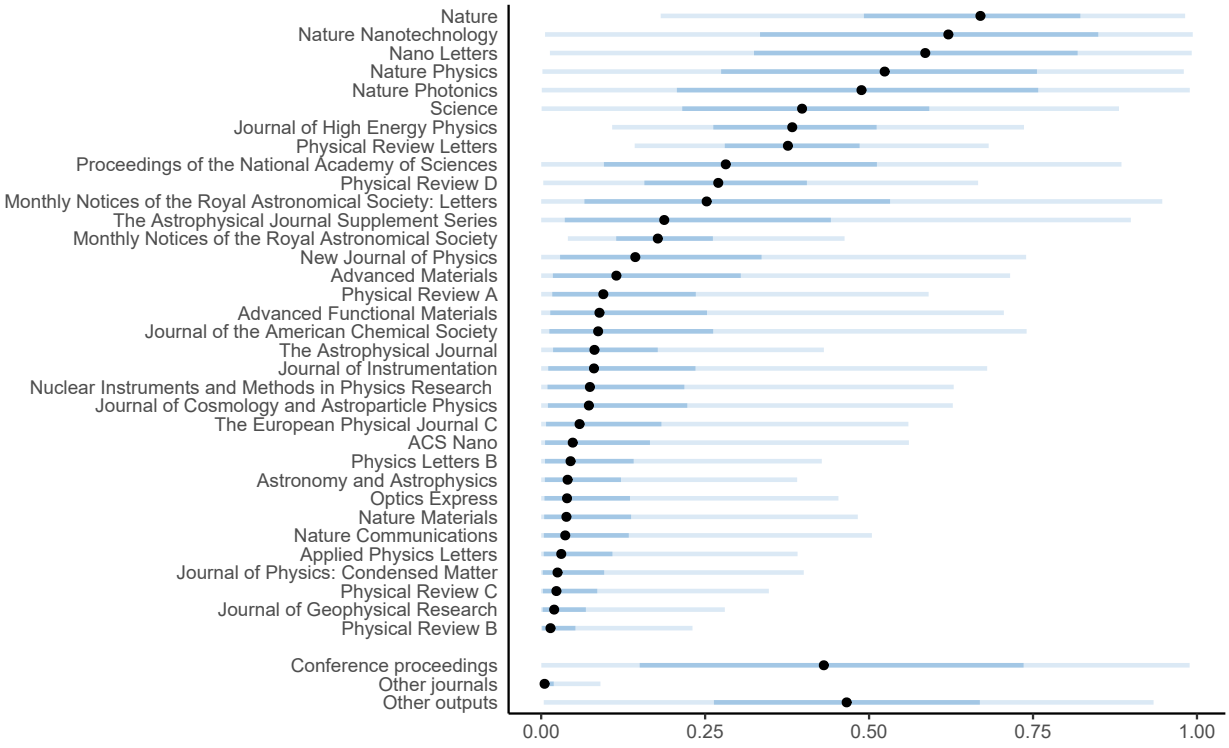
Figure 6.12 provides league tables of estimated journal REF success probabilities in Chemistry. This field, unlike the others studied here, seems to be dominated by popular general science outlets, in *PNAS*, *Nature* and *Science*, rather than dedicated chemistry journals. There may be some dependence on types of articles published: some periodicals print different mixtures of 'full' research papers and communications (letters). *Nature Chemistry* and *Nature Communications* fall lower in the ranking than might be expected (Bugg, 2019; Scott, 2019). This appears to be simply a result of work published in journals being submitted by several low-scoring institutions that were not awarded many 4* ratings in the REF.

The expected versus actual institutional results are plotted in Figure 6.13. The pattern around the line of $y = x$ is similar to that in the other sub-panels: broadly a linear relationship, but with lower-scoring institutions having higher predicted than actual results, and the converse for stronger institutions. There are no noticeable outliers.

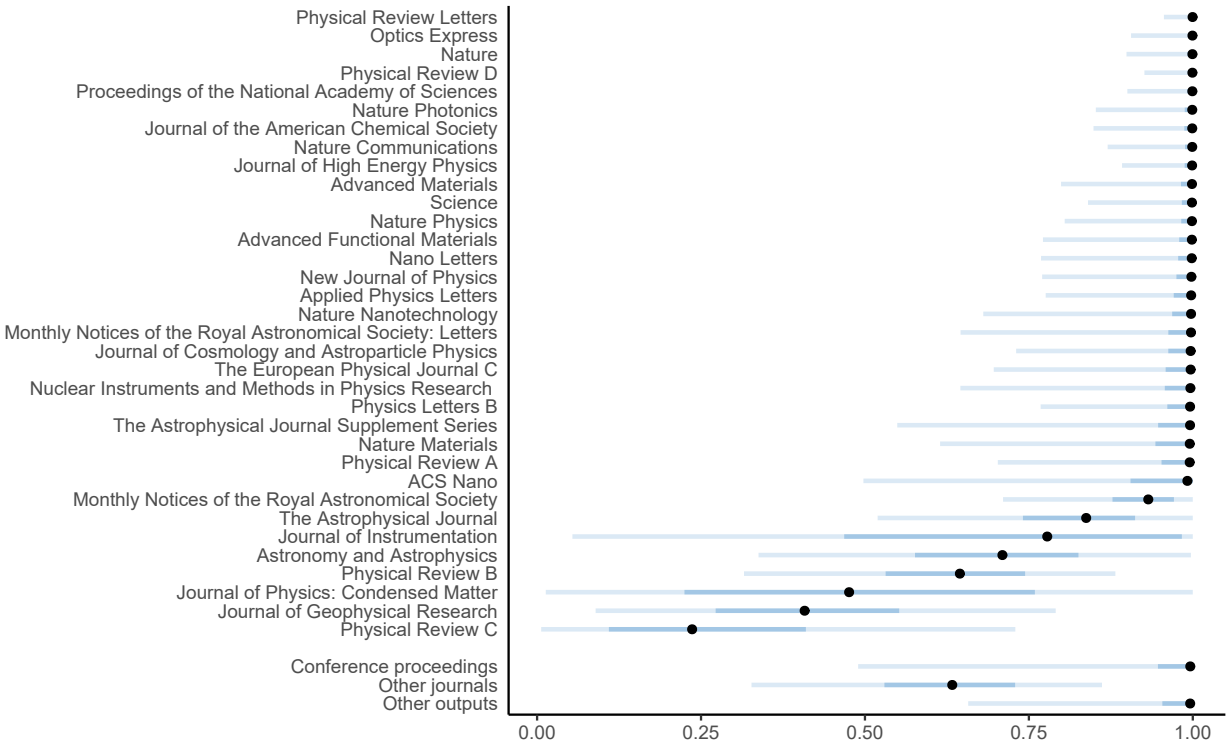
By summary measures, the median index of similarity in Chemistry is 8.2% and the median proportion of research funding that would need to be reallocated would be 5.6%. The posterior distributions are plotted in Figure 6.5. Performance, according to these metrics, appears similar to for other fields.

6.6.5 Comparison with Journal Impact Factors

Using data from Clarivate Analytics' *Journal Citation Reports*, we can compare the latent journal REF effects with journal impact factors for the respective year. For this article, we use the 2014 edition of

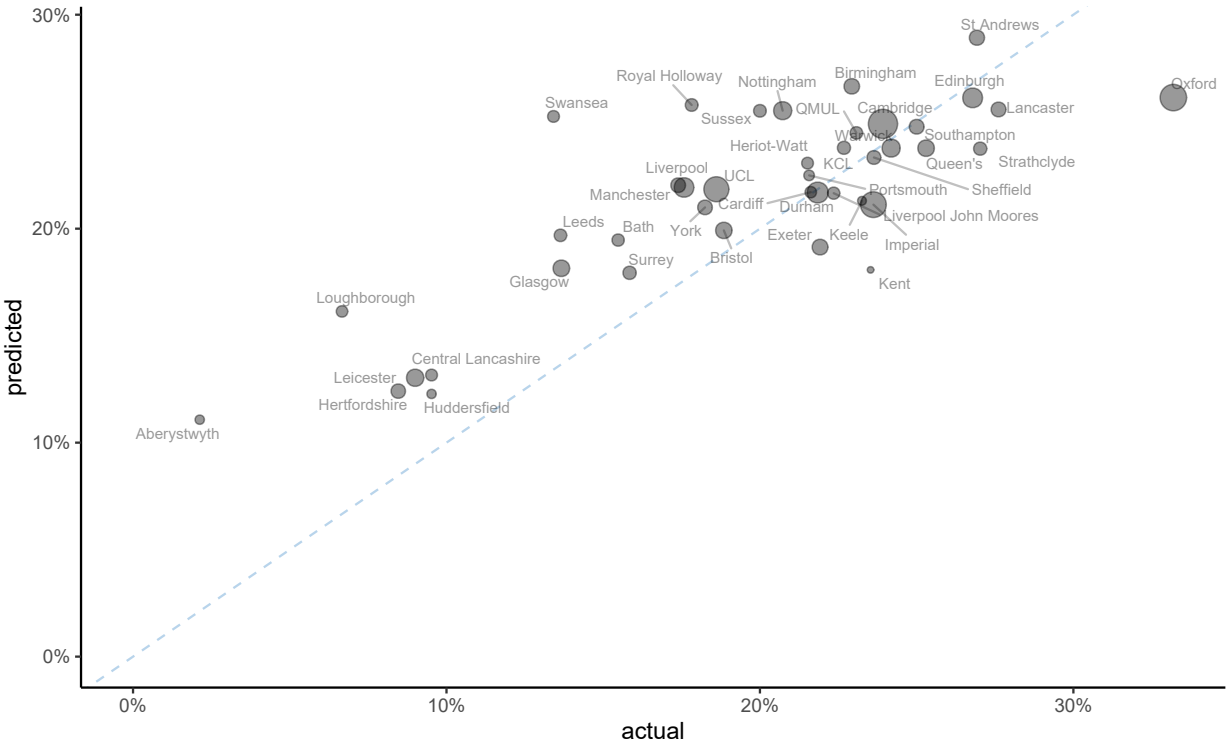


(a) Probability of 4*

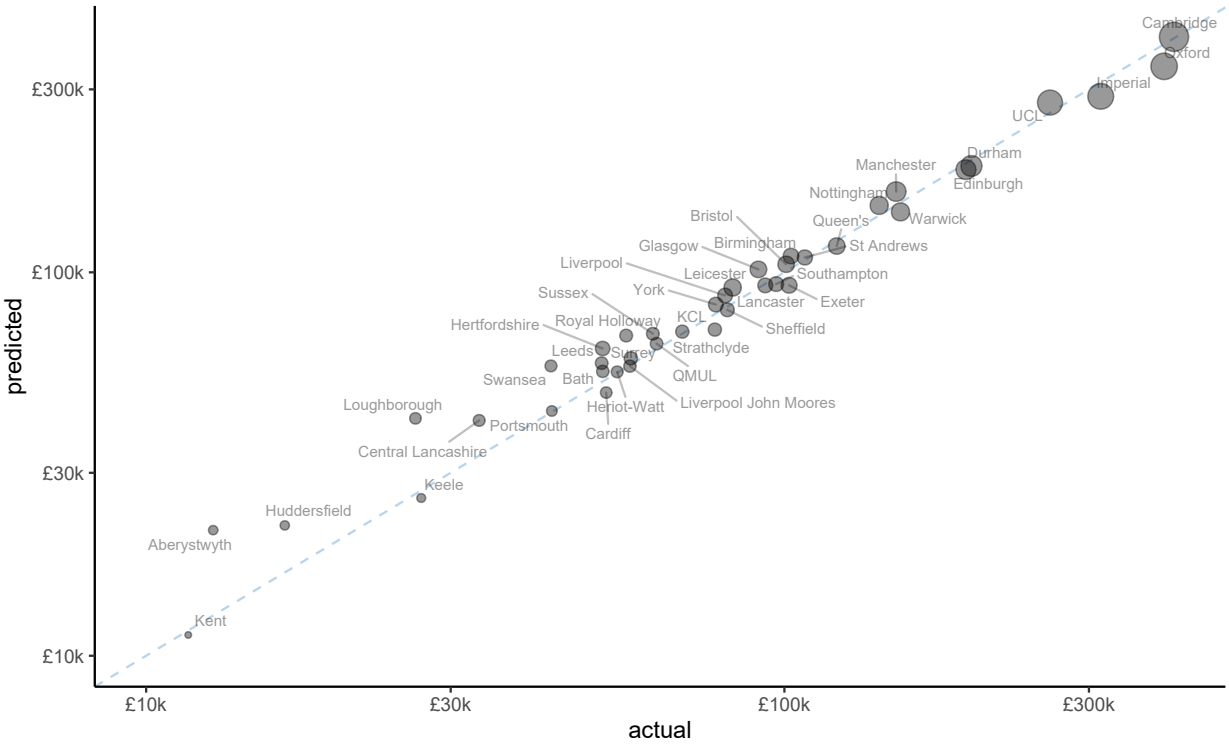


(b) Probability of 3* or 4*

Figure 6.10: Median estimated journal success probabilities of 4* ratings in Physics. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014

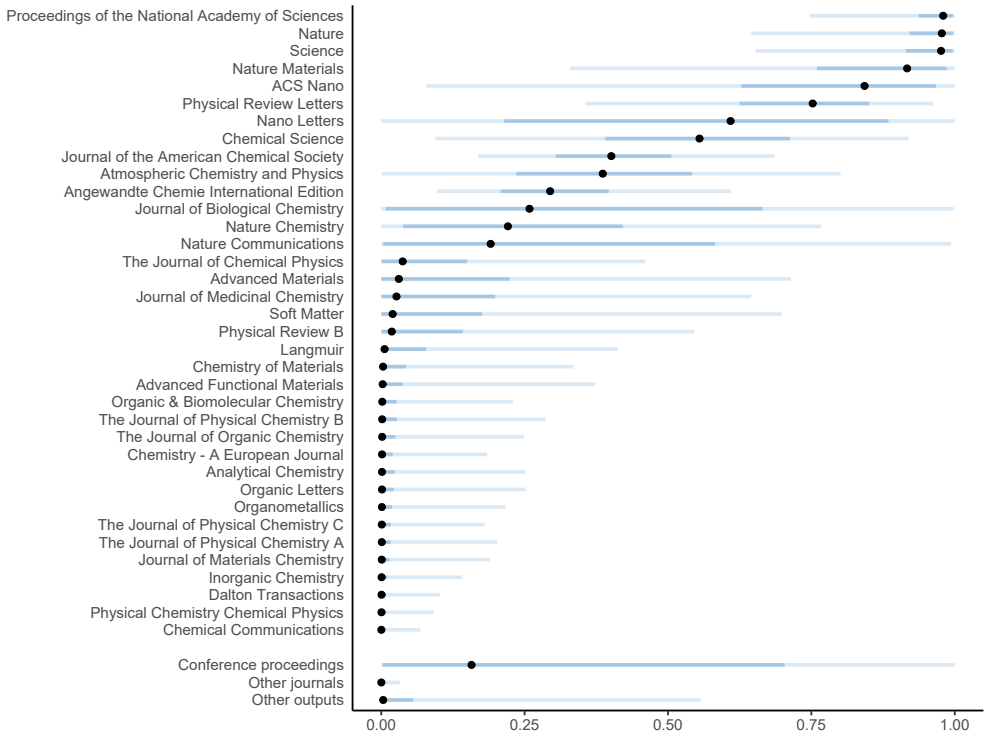


(a) % articles at 4*

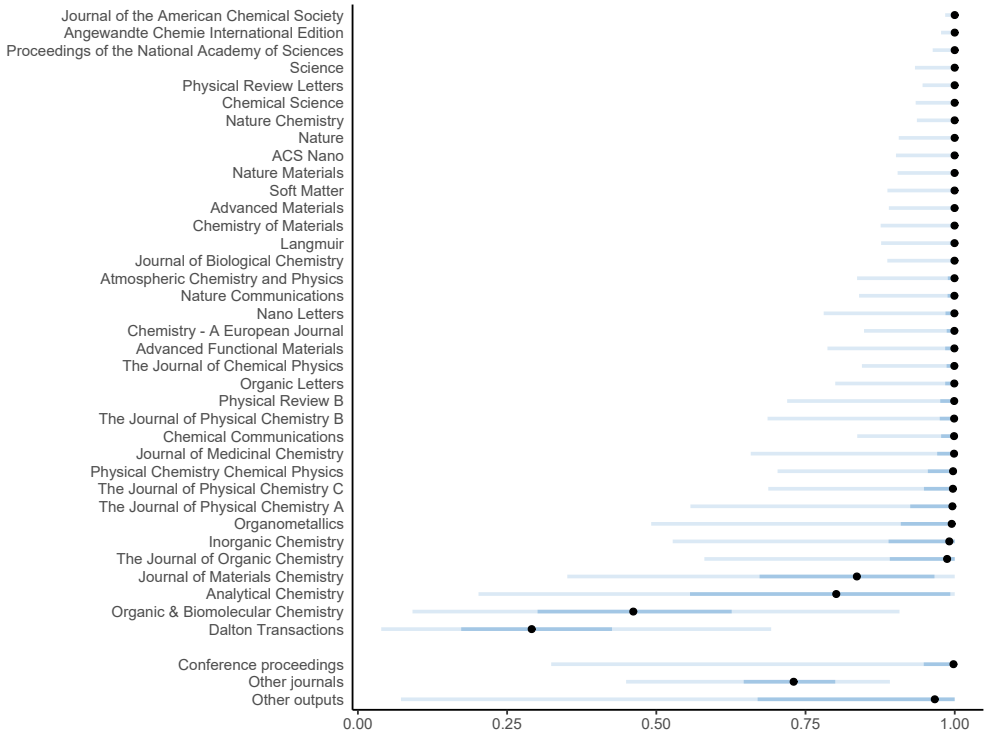


(b) funding allocation

Figure 6.11: Predictions versus observed REF2014 results for institutions submitting outputs to the Physics subpanel, with point sizes proportional to number of FTE staff

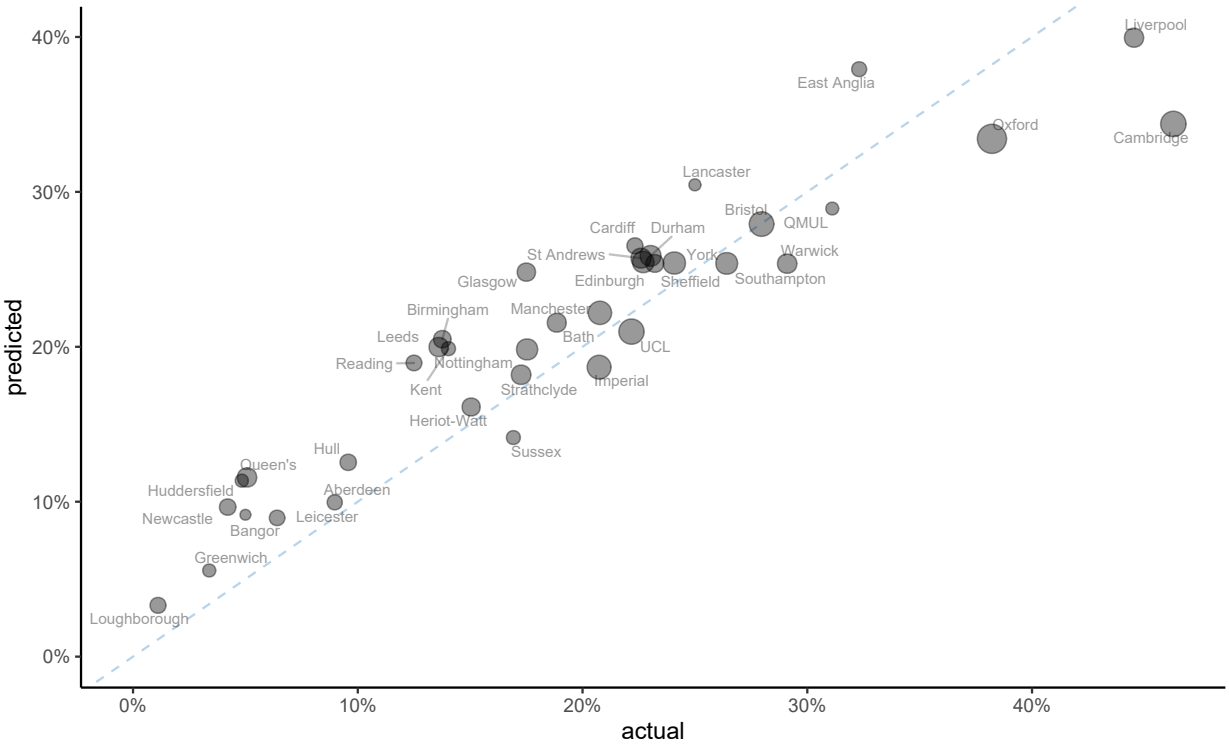


(a) Probability of 4*

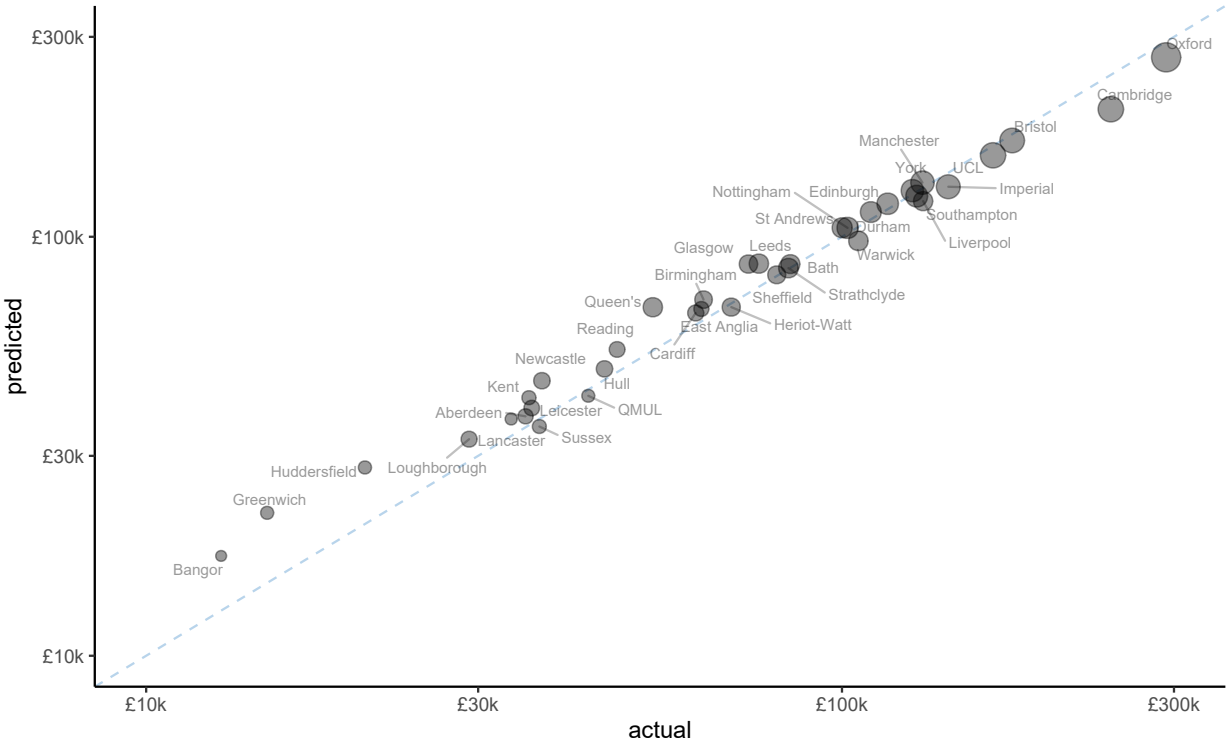


(b) Probability of 3* or 4*

Figure 6.12: Median estimated journal success probabilities of 4* ratings in Chemistry. Shaded line segments represent 50% and 95% posterior intervals. Named journals had 30 or more articles submitted in REF2014



(a) % articles at 4*



(b) funding allocation

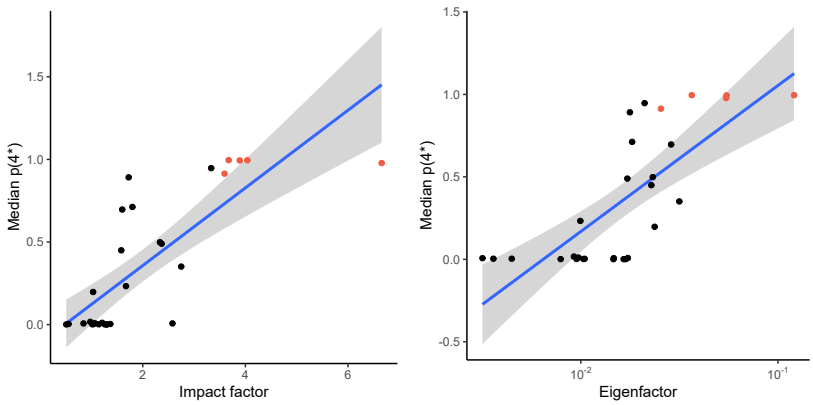
Figure 6.13: Predictions versus observed REF2014 results for institutions submitting outputs to the Chemistry sub-panel, with point sizes proportional to number of FTE staff

Journal Citation Reports, as this is based on citation data from the preceding two years. (One could also consider the 2013 edition, though the results should not be too different.)

It may also be possible to compare with rival metrics, such as the CiteScore and Scimago Journal Rank (SJR), Scopus's versions of the impact factor and the Eigenfactor, respectively, however we do not make those comparisons here.

Economics & Econometrics

Comparisons are plotted in Figure 6.14. Note the logarithmic scale for the Eigenfactor score. Broadly speaking, there is a (weak) positive correlation between both citation metrics and the probability of attaining 4* in the REF. Evidence for the supposed dominance of the 'top 5' economics journals is mixed. Whilst these periodicals are indeed highly ranked by journal impact factor, Eigenfactor and apparent REF effect, they do not completely dominate the top five spots, so their reputation must depend on other factors or perhaps be undeserved. Moreover, as economists have explicitly known of the 'top 5' designation for years, it may be a self-fulfilling prophecy.



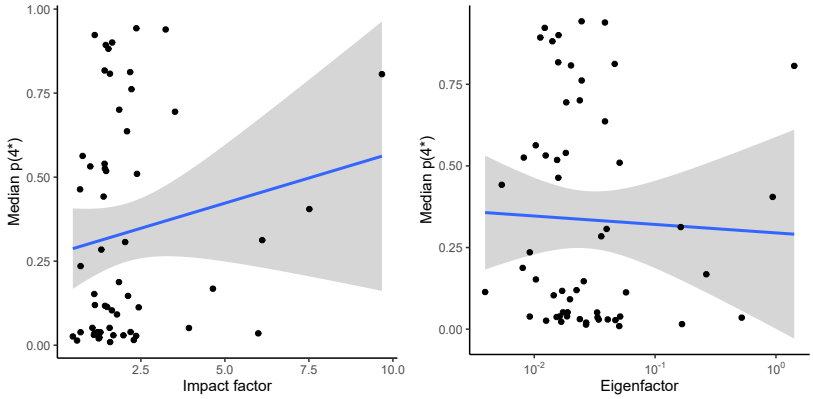
(a) Journal impact factor (2014)

(b) Eigenfactor (2014)

Figure 6.14: Comparison of Economics and Econometrics journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit. So-called 'top 5' journals are highlighted in red

Mathematical Sciences

In Mathematical Sciences, however, there is almost no correlation between journal impact factor and the estimated probability of 4* in the REF; see Figure 6.15. This phenomenon could partly be explained by mathematical journals generally receiving lower impact factors; mathematics papers tend to have short reference lists and take longer to be noticed, when compared with publications in microbiology and other applied disciplines, so the journal impact factor (roughly speaking, counting citations over two years) is an especially poor metric for mathematics work. Most mathematics journals here had an impact factor of around 1 or 2, so most of the variation between those scores might be attributed to random noise—see the left hand side of Figure 6.15a.

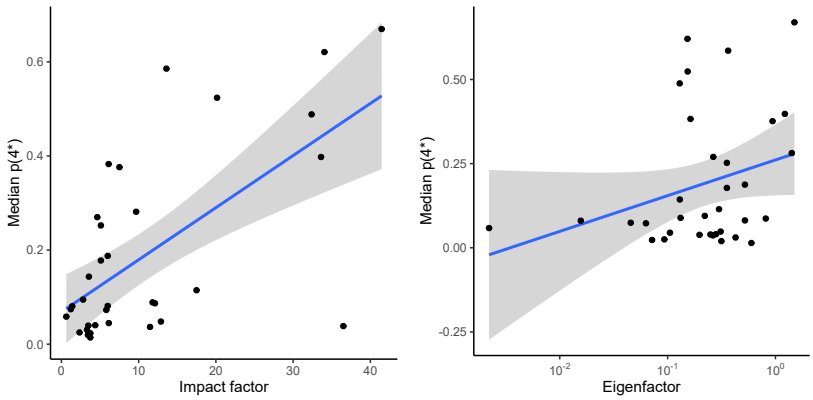


(a) Journal impact factor (2014) (b) Eigenfactor (2014)

Figure 6.15: Comparison of Mathematical Sciences journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit

Physics

A positive correlation is present between Clarivate citation metrics and estimated probability of 4* for Physics as illustrated in Figure 6.16.



(a) Journal impact factor (2014) (b) Eigenfactor (2014)

Figure 6.16: Comparison of Physics journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit

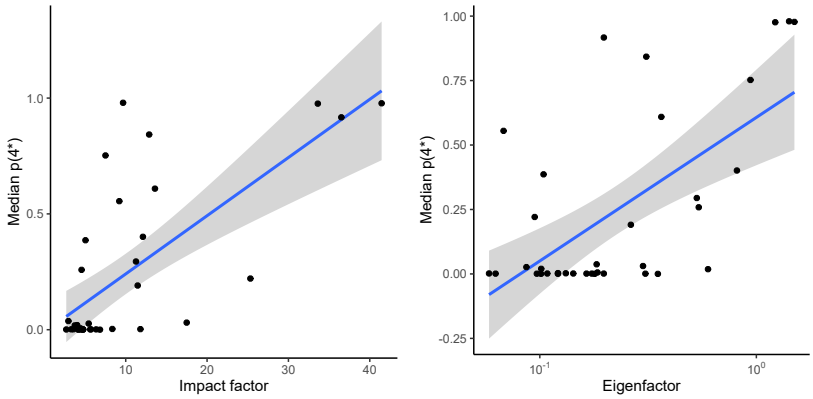
Chemistry

The story for Chemistry is hard to interpret because so many journals have median estimated 4* probabilities close to zero. But the top three journals by impact factor were also estimated to have the highest chances of their articles attaining 4* in the REF. See Figure 6.17.

6.7 Discussion

This chapter has explored the relationship between published REF2014 results and the journals in which institutions published research outputs submitted for REF2014 assessment. The results are informative in various ways, including:

- implied rankings of the main journals from which work was submitted in each REF2014 sub-panel, together with measures of uncertainty on such rankings; and



(a) Journal impact factor (2014)

(b) Eigenfactor (2014)

Figure 6.17: Comparison of Chemistry journals' estimated probabilities of attaining 4* in the REF, versus Clarivate journal citation metrics, with line of best fit

- for each REF2014 sub-panel studied, measurement of the *maximum* extent to which REF2014 outcomes can be explained (retrospectively) by the identities of the journals from which work was submitted by each institution.

One reassuring aspect of the journal rankings derived for the four disciplines studied here is that they broadly agree with the informed opinions (informally elicited) of senior Warwick academics in those disciplines. That is to say, for the main 'named' journals in each field, the estimates and uncertainty intervals for the journals' probabilities of attaining 4* ratings in REF2014 made sense in the minds of the experts who were consulted. Had the opposite been found, it would have been a strong reason to distrust the statistical methodology used here.

Our analysis of four disciplines found that in each of them there is—as expected—a strong or very strong relationship between the composition of journals seen in an institution's REF2014 submission and its published REF2014 Outputs profile results.

Naïvely, one might infer that the REF could therefore be replaced—at least for some disciplines—by a more automated 'algorithmic' assessment that assigns quality ratings based on the journal in which each piece of research is published, rather than on an expert panel's reading of the work itself. However, such an interpretation would not be justified. As well as the potential for such an algorithmic approach to produce undesirable changes in behaviour, it is important to emphasise two aspects of our analysis. Specifically:

1. The analysis performed here is *retrospective*, not predictive. The question asked, in each discipline, was effectively: if we imagine that the REF2014 panel based its assessments on journal identities alone, then what set of 'journal quality' scores would yield the best match with the actual published REF2014 results? How good would such a 'best match' be? The implied 'journal quality' scores in our analysis came directly from the REF2014 *results*; they were not known in advance by the REF panel, nor were they based on any explanatory covariates other than the journal identities themselves.

2. Although strong correlation was found between REF outcomes and aggregated 'journal quality' scores (see Figures 6.4, 6.9, 6.11 and 6.13), there is a clear *pattern* of deviation from that relationship, for each of the disciplines studied here. The 'top' institutions are seen typically to do better in REF2014 than their aggregated 'journal quality' scores would suggest; and conversely institutions at the other end of the scale tend to do worse, relative to purely journal-based scores. This indicates that REF assessment panels are in fact doing more than simply using journal identity to determine research quality. This finding is unsurprising: the published remit of REF panels is to *read* the submitted research and evaluate its quality against clearly stated criteria. With that in mind, it is fully to be expected that a diligent REF panel will distinguish the 'best' papers in each journal from those papers that are more ordinary.

It could perhaps still be argued that the relationship between journal-based scores and REF outcomes is sufficiently strong that deviations from it could be ignored, in the interest of reducing the overall cost of the REF exercise. But the clear *pattern* of deviation described in point 2 implies that the resulting redistribution (of research funding, but also prestige) would systematically disadvantage those institutions where predominantly top-quality research is done. While such redistribution of funds might represent a fairly modest fraction of the national funding total, its effects would systematically be concentrated in a few institutions at opposite ends of the scale.

Furthermore, the notion of judging work based on the container in which it is published, rather than on its own merits, seems to miss the point of research assessment entirely. As Traag and Waltman (2019) points out, by relying on metrics, even those which correlate strongly with peer review results, 'the goal of fostering "high quality" science may become displaced by the goal of obtaining a high metric' and have unintended consequences such as 'favouring problematic research methods'.

More pragmatically, there is nothing to say that the esteem of academic journals in 2014 will remain constant until 2021. Editors and authors change and publications can go defunct or start anew in such a long period. Mryglod et al. (2015a; 2015b) already showed that one research assessment exercise cannot necessarily be used to predict the next.

Perhaps the most interesting avenue for future research would be to apply these methods to *all* subject areas in the REF and determine which fields are most beholden to the effect of journals on institutional rankings. Data for all 36 units of assessment in the REF are readily available, and it should be straightforward to apply the methods developed here to those other fields. With the 2021 REF approaching, this could be a topic of interest to many in academia, publishing and research assessment. As seen in Table 6.2, however,

subjects in the hard sciences tend to submit to scholarly journals more than other fields, such as the arts, who may produce books or artefacts, so the methodology would need to be adapted carefully for such areas, if indeed it can be applied at all.

7
Concluding remarks

Citation analysis is a controversial topic and mostly centres around heuristic methods to estimate a latent variable quantifying 'quality' or 'prestige' in different contexts. Standard or established practices in this area—particularly when (mis-)used for research assessment—have been found wanting, through methodological problems, lack of transparency and negative effects on research practices. The heterogeneity of academia makes metrics and rankings difficult to compare between subject areas, the delineations of which are also often subjective and arbitrary.

A tricky challenge then: to what extent is it feasible to measure something that is ill-defined, from discrepant data sources, produce a seemingly—or at least defensibly—'objective' metric and make it simple enough for non-technical end-users, such as librarians and administrators, to understand? At the very least, we can make a valiant attempt in highlighting and nudging the industry towards 'least worst' practices.

In this thesis, we have discussed the pitfalls of using measures such as the journal impact factor to influence decisions in research and research administration. While some other metrics try to mitigate its well-documented flaws, most can not be considered (by statisticians) to be 'statistical' without some quantification of uncertainty, and many assume a static academy where subject areas are immutable and citations all carry equal weight.

Building on the work of Varin et al. (2016), we pitched the palatable-to-statisticians generalized linear model of Stigler (1994) against the Eigenfactor metrics of Bergstrom et al. (2008). In so doing, we uncovered a deep theoretical connection between these otherwise seemingly distinct ranking systems, with the potential to motivate the development of 'error bars' and residual diagnostics for descriptive statistics like PageRank that otherwise lacked them.

In practical terms we revive the 40-year-old influence weight metric of Pinski and Narin (1976), or 'Scroogefactor', that can be computed as quickly and easily as PageRank and offers tangible benefits that might discourage citation manipulation and help control for size biases in the analysis of journal networks. An empirical case study showed that this metric—like the Stigler model—can identify the statistical journals considered 'best' by academic statis-

ticians.

An ad-hoc analysis considering a hand-picked group of statistics journals is not necessarily scalable or replicable for different fields in the wider academic community, however. Diversity in publishing practices between academic disciplines reduces the utility of a single overall ranking (such as the impact factor) without some way of grouping publications by subject area. One place to start a field-by-field analysis would be to use published lists of subject categories such as those in the *Web of Science*, but these are not transparent and often inconsistent.

With so much research funding—often allocated at the university department level—at stake, it would be prudent for field definitions, like citation metrics, to be transparent, reproducible and data-driven. Some standard algorithmic methods for discovering communities in networks may be applied to citation data, but those accepted as 'best practice' have problems, in theory and in application, on large, directed, weighted networks. We had reasonable success using the Infomap algorithm, and for future studies recommend the use of community residuals as a structured framework for evaluating the quality of a clustering, as well as for identifying possible anomalous publishing behaviour, such as citation cartels.

We extended our ranking methodology to measure influence *between* as well as *within* disciplines, by a simple extension and reapplication of the Stigler model, aggregating journals into field-representative super-journals.

Among subject categories (both from the *Web of Science* and algorithmically-generated) we notice a prevailing flow of influence from biological sciences, medicine, statistics and the social sciences to the more fundamental fields such as mathematics and the physical sciences, as well as engineering. We posit this behaviour may be descriptive of routine behaviour, where (the introductions to) academic papers are often written with one eye on 'impact' and relevance to applications. Additionally, theoretical contributions to the literature often take longer to achieve recognition and this is not necessarily represented through citations. In a cruel twist of fate, our inter-field bibliometric analysis finds *bibliometrics* to be one of the least influential disciplines of all, when measured in this way.

The use of an 'other fields' super-journal, representing external citation exchange in a within-field ranking, is a novel way to measure the influence of an academic discipline, and the journals within it, in a wider context. Directly comparing intra-field with wider influence rankings gives an insight into the level of insularity or interdisciplinarity of different groups of publications.

In order for our proposed ranking methods and residual diagnostics to be applied in the real world, we have demonstrated several easy ways by which it is possible to obtain citation data reliably, without taking the standard Web of Science or Scopus databases for granted. Ideally, this should open up bibliometrics to non-specialists who may not have already built privileged relation-

ships with data providers such as Elsevier or Clarivate. Moreover, a diversity of data sources should allow verification of results, and discourage conclusions that cannot be replicated. It also allows future studies at the author or institution level rather than merely comparing journals, which were our main focus throughout this thesis.

Anecdotally, journal prestige plays a big role in academic life, apparently influencing hiring, promotion and funding decisions. In Chapter 6, we examined the relationship between academic journals and research assessment/funding in the United Kingdom, based on submissions and results data from the 2014 Research Excellence Framework. We presented frequentist and Bayesian approaches to this missing data problem, applied on a larger scale than previous studies on the REF. The `ref2014` package provides convenient ways by which these analyses may be replicated.

The analysis found a strong relationship between journal submissions and published REF results. It also provided a set of journal rankings based on the estimated probability of publications receiving certain quality scores. These implied rankings mostly conformed with informal expert judgements of the journals' reputations in each discipline. However, automating the research assessment process with such a methodology would not be justified, not just because the model is retrospective, but because using journal identities alone to produce quality assessments would disadvantage top-performing universities, as well as introducing perverse incentives for researchers.

To conclude, we have reviewed the current state of citation-based assessment and found it wanting. Through some relatively simple extensions of existing techniques and reapplications of old ideas, it may be possible to overcome some of the shortfalls of proxy measures that are in use today. Modelling the relationships that research publications have—with each other and with the academic community—offers the opportunity to learn things about academic behaviour that are not captured in simple lists of average citation counts.

Bibliography

Adler, R., Ewing, J., and Taylor, P. (2008). *Citation statistics: A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)*. Joint Committee on Quantitative Assessment of Research.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 3rd edition.

Ahmed, A. and Xing, E. P. (2010). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010)*.

Amin, M. and Mabe, M. A. (2003). Impact factors: use and abuse. *Medicina (Buenos Aires)*, 63(4):347–354.

Anderson, C. J., Wasserman, S., and Faust, K. (1992). Building stochastic blockmodels. *Social Networks*, 14(1-2):137–161.

Aria, M. and Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4):959–975.

Arnold, D. N. and Fowler, K. K. (2011). Nefarious numbers. *Notices of the AMS*, 58(3):434–437.

Avrachenkov, K., Ribeiro, B., and Towsley, D. (2010). Improving random walk estimation accuracy with uniform restarts. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 98–109. Springer.

Baccini, A. and Nicolao, G. D. (2016). Do they agree? bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, 108(3):1651–1671.

Baker, D. (2012). Google h5 vs Thomson impact factor. `https://bakerdh.wordpress.com/2012/04/05/google-h5-vs-thomson-impact-factor/`. Accessed 2020-02-13.

Balbuena, L. D. (2018). The UK Research Excellence Framework and the Matthew effect: Insights from machine learning. *PLOS ONE*, 13(11):e0207919.

Ball, R. C. (2019). Personal communication.

Bartsch, K. (2017). The Napster moment: Access and innovation in academic publishing. *Information Services & Use*, 37(3):343–348.

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, 489(7415):179–179.

Beall, J. (2017). What I learned from predatory publishers. *Biochemia Medica*, 27(2):273–278.

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splittgerber, R., Stephenson, J., Tower, C., Walton, R. G., and Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education*, 31(2):145–152.

Berenbaum, M. R. (2019). Impact factor impacts on early-career scientist careers. *Proceedings of the National Academy of Sciences*, 116(34):16659–16662.

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68(5):314–316.

Bergstrom, C. T. and West, J. D. (2016). Comparing Impact Factor and Scopus CiteScore. `http://eigenfactor.org/projects/posts/citescore.php`. Accessed 2018-04-27.

Bergstrom, C. T., West, J. D., and Wiseman, M. A. (2008). The Eigenfactor metrics. *The Journal of Neuroscience*, 28(45):11433–11434.

Bernal, J. D. (1939). *The Social Function of Science*. George Routledge & Sons Ltd.

Biswas, A. and Biswas, B. (2016). A framework for analyzing community detection algorithms. In *Technology Symposium (TechSym), 2016 IEEE Students'*, pages 61–66. IEEE.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM Press.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bof, N., Baggio, G., and Zampieri, S. (2017). On the role of network centrality in the controllability of complex networks. *IEEE Transactions on Control of Network Systems*, 4(3):643–653.

Bohannon, J. (2013). Who's afraid of peer review? *Science*, 342(6154):60–65.

Bohannon, J. (2015). Updated: Editor quits journal over pay-for-expedited peer-review offer. *Science*. Accessed 2018-05-03.

Bohannon, J. (2016). Who's downloading pirated papers? everyone. *Science*, 352(6285):508–512.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3-4):324–345.

Brase, J. (2009). DataCite - a global registration agency for research data. In *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*. IEEE.

Brembs, B. (2016). Sci-Hub as necessary, effective civil disobedience. `http://openscience.ens.fr/OTHER/SCIHUB/2016_02_25_Bjorn_Brembs_about_SciHub.pdf`. Accessed 2018-05-08.

Broderick, T., Pitman, J., and Jordan, M. I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836.

Brown, P. J. and Payne, C. D. (1986). Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association*, 81(394):452–460.

Bugg, T. (2019). Personal communication.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Caussinus, H. (1965). Contribution a l'analyse statistique des tableaux de correlation. In *Annales de la Faculté des Sciences de Toulouse*, volume 29, pages 77–183. Université Paul Sabatier.

Caves, C. M. (2014). High-impact-factor syndrome. *APS News*, 23(10):8–9.

Chamberlain, S. (2018). *microdemic: 'Microsoft Academic' API Client*. R package version 0.4.0.

Chamberlain, S. (2019). *citecorp: Client for the Open Citations Corpus*. R package version 0.2.2.

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., and Ram, K. (2019). *rcrossref: Client for Various 'CrossRef' APIs*. R package version 0.9.2.

Chawla, D. S. (2018). The undercover academic keeping tabs on 'predatory' publishing. *Nature*, 555(7697):422–423.

Cho, W. K. T. (1998). Iff the assumption fits...: a comment on the King ecological inference solution. *Political Analysis*, 7:143–163.

Clarivate Analytics (2019). Journal Citation Reports. https://jcr.clarivate.com.

Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.

Cochran, A. (2016). A funny thing happened on the way to OA. https://scholarlykitchen.sspnet.org/2016/02/25/a-funny-thing-happened-on-the-way-to-oa/. Accessed 2018-05-08.

Coldewey, D. (2018). Thousands of academics spurn Nature's new paid-access machine learning journal. https://techcrunch.com/2018/05/01/thousands-of-academics-spurn-natures-new-paid-access-machine-learning-journal/. Accessed 2018-05-08.

Colquhoun, D. and Plested, A. (2014). Why you should ignore altmetrics and other bibliometric nightmares. http://www.dcscience.net/2014/01/16/why-you-should-ignore-altmetrics-and-other-bibliometric-nightmares/. Accessed 2018-04-26.

Creusefond, J., Largillier, T., and Peyronnet, S. (2016). On the evaluation potential of quality functions in community detection for different contexts. In Wierzbicki, A., Brandes, U., Schweitzer, F., and Pedreschi, D., editors, *Advances in Network Science: 12th International Conference and School, NetSci-X 2016*, Lecture Notes in Computer Science. Springer.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

da Silva, J. A. T. and Dobránszki, J. (2014). Problems with traditional science publishing and finding a wider niche for post-publication peer review. *Accountability in Research*, 22(1):22–40.

da Silva, J. A. T. and Memon, A. R. (2017). CiteScore: A cite for sore eyes, or a valuable, transparent metric? *Scientometrics*, 111(1):553–556.

Davis, P. (2016). CiteScore–flawed but still a game changer. https://scholarlykitchen.sspnet.org/2016/12/12/citescore-flawed-but-still-a-game-changer/. Accessed 2018-04-27.

Davis, P. (2017). Scientific Reports overtakes PLoS One as largest megajournal. https://scholarlykitchen.sspnet.org/2017/04/06/

scientific-reports-overtakes-plos-one-as-largest-megajournal/. Accessed 2018-05-8.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Dias, E. S., Castonguay, D., and Dourado, M. C. (2016). Algorithms and properties for positive symmetrizable matrices. *TEMA (São Carlos)*, 17(2):187.

Dickens, C. (1843). *A Christmas Carol*. Chapman & Hall.

Duane, S., Kennedy, A., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222.

Dubey, A., Hefny, A., Williamson, S., and Xing, E. P. (2013). A nonparametric mixture model for topic modeling over time. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics.

Duncan, O. D. and Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18(6):665.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210.

Eddelbuettel, D. and Balamuta, J. J. (2017). Extending R with C++: A Brief Introduction to Rcpp. *PeerJ Preprints*, 5:e3188v1.

Edwards, S. F. and Anderson, P. W. (1975). Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965–974.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, pages 1–26.

Eldardiry, H. and Neville, J. (2008). A resampling technique for relational data graphs. In *Proceedings of the 2nd SNA workshop, 14th ACM SIGKDD conference on knowledge discovery and data mining*.

Else, H. (2015). Research funding formula tweaked after REF 2014 results. https://www.timeshighereducation.com/news/research-funding-formula-tweaked-after-ref-2014-results/2018685.article. Accessed 2019-04-09.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. John Wiley and Sons Ltd, 5th edition.

Falagas, M. E. and Alexiou, V. G. (2008). The top-ten in journal impact factor manipulation. *Archivum Immunologiae et Therapiae Experimentalis*, 56(4):223–226.

Firth, D. (2015). *qvcalc: Quasi Variances for Factor Effects in Statistical Models*. R package version 0.8-9.

Firth, D. and de Menezes, R. X. (2004). Quasi-variances. *Biometrika*, 91(1):65–80.

Firth, D. and Turner, H. L. (2012). Bradley–Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software*, 48(9).

Flaxman, S., Sutherland, D., Wang, Y.-X., and Teh, Y. W. (2016). Understanding the 2016 US presidential election using ecological inference and distribution regression with census microdata.

Flaxman, S. R., Wang, Y.-X., and Smola, A. J. (2015). Who supported Obama in 2012? In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. ACM Press.

Fleck, C. (2013). The impact factor fetishism. *European Journal of Sociology*, 54(2):327–356.

Fog, A. (2015). *BiasedUrn: biased urn model distributions*. R package version 1.07.

Fortunato, S. (2007). Quality functions in community detection. In Kertész, J., Bornholdt, S., and Mantegna, R. N., editors, *Noise and Stochastics in Complex Systems and Finance*. SPIE.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3):75–174.

Fortunato, S. and Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A.-L. (2018). Science of science. *Science*, 359(6379):eaao0185.

Franceschet, M. (2010). Ten good reasons to use the Eigenfactor metrics. *Information Processing & Management*, 46(5):555–558.

Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM*, 54(6):92–101.

Franceschet, M. and Costantini, A. (2011). The first italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, 5(2):275–291.

Freedman, D. A., Ostland, M., Roberts, M. R., and Klein, S. P. (1999). Reply to G. King. *Journal of the American Statistical Association*, 94(445):355–257.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178:471–479.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, 1(4):359–375.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295(1):90.

Geller, N. L. (1978). On the citation influence methodology of Pinski and Narin. *Information Processing & Management*, 14(2):93–95.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Gleich, D. F. (2015). PageRank beyond the Web. *SIAM Review*, 57(3):321–363.

Golub, G. H. and Meyer, Jr, C. D. (1986). Using the qr factorization and group inversion to compute, differentiate, and estimate the sensitivity of stationary probabilities for markov chains. *SIAM Journal on Algebraic Discrete Methods*, 7(2):273–281.

González-Pereira, B., Guerrero-Bote, V. P., and Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3):379–391.

Good, B. H., de Montjoye, Y.-A., and Clauset, A. (2010). The performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106.

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review*, 18(6):663.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, 64(6):610–625.

Gregory, S. (2008). A fast algorithm to find overlapping communities in networks. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 408–423. Springer Nature.

Gregory, S. (2009). Finding overlapping communities using disjoint community detection algorithms. In *Complex Networks*, pages 47–61. Springer.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5228–5235.

Groen-Xu, M., Teixeira, P., Voigt, T., and Knapp, B. (2017). Short-termism in science: Evidence from the UK research excellence framework. *SSRN Electronic Journal*.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in R. *Bioinformatics*, 30(19):2811–2812.

Harnad, S. (1994). Publicly retrievable FTP archives for esoteric science and scholarship: a subversive proposal. `https://groups.google.com/d/msg/bit.listserv.vpiej-l/BoKENhK0_00/2MF9QBO9s2IJ`. Accessed 2018-05-04.

Harnad, S. (2014). Crowd-sourced peer review: Substitute or supplement for the current outdated system? http://blogs.lse.ac.uk/impactofsocialsciences/2014/08/21/crowd-sourced-peer-review-substitute-or-supplement/. Accessed 2018-05-02.

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., and Hilf, E. R. (2004). The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4):310–314.

Harzing, A.-W. (2016). Microsoft Academic (Search): a phoenix arisen from the ashes? *Scientometrics*, 108(3):1637–1647.

Harzing, A.-W. and Alakangas, S. (2017). Microsoft Academic is one year old: the phoenix is ready to leave the nest. *Scientometrics*, 112(3):1887–1894.

Hastings, M. B. (2006). Community detection as an inference problem. *Physical Review E*, 74(3).

Heckman, J. and Moktan, S. (2018). Publishing and promotion in economics: The tyranny of the top five. techreport, National Bureau of Economic Research. https://www.nber.org/papers/w25093.

Heimo, T., Kumpula, J. M., Kaski, K., and Saramäki, J. (2008). Detecting modules in dense weighted networks with the potts method. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08007.

Hill, S. (2017). A paper exploring REF, funding and journal metrics. http://stevenhill.org.uk/a-paper-exploring-REF-funding-and-journal-metrics/. Accessed 2019-09-10.

Himmelstein, D. S., Romero, A. R., Levernier, J. G., Munro, T. A., McLaughlin, S. R., Tzovaras, B. G., and Greene, C. S. (2018). Sci-Hub provides access to nearly all scholarly literature. *eLife*, 7.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Hofman, J. M. and Wiggins, C. H. (2008). Bayesian approach to network modularity. *Physical Review Letters*, 100(25).

Holden, G., Rosenberg, G., and Barker, K. (2005). Bibliometrics. *Social Work in Health Care*, 41(3-4):67–92.

Holden, G., Rosenberg, G., Barker, K., and Onghena, P. (2006). Should decisions about your hiring, reappointment, tenure, or promotion use the impact factor score as a proxy indicator of the impact of your scholarship? *MedGenMed : Medscape general medicine*, 8:21.

Hole, A. R. (2017). Ranking economics journals using data from a national research evaluation exercise. *Oxford Bulletin of Economics and Statistics*, 79(5):621–636.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, 59:41–51.

Hug, S. E., Ochsner, M., and Brändle, M. P. (2017). Citation analysis with Microsoft Academic. *Scientometrics*, 111(1):371–378.

I4OC (2020). I4oc: Initiative for open citations. `https://i4oc.org/`. Accessed 2020-01-08.

Jannarone, R. J., Yu, K. F., and Laughlin, J. E. (1990). Easy Bayes estimation for Rasch-type models. *Psychometrika*, 55(3):449–460.

Jump, P. (2011). Nature's open-access offering may sound death knell for subs model. `https://www.timeshighereducation.com/news/natures-open-access-offering-may-sound-death-knell-for-subs-model/414822.article`. Accessed 2018-05-04.

Jump, P. (2013). 'Game' of one-fifth? part-time contracts rise in run-up to REF. *Times Higher Education*, (2120):6.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1).

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kaye, E. and Firth, D. (2017). *BradleyTerryScalable: Fits the Bradley-Terry Model to Potentially Large and Sparse Networks of Comparison Data*. R package version 0.1.0.

Keirstead, J. (2016). *scholar: analyse citation data from Google Scholar*. R package version 0.1.5.

Kelly, F. P. (1979). *Reversibility and stochastic networks*. Wiley.

King, G. (1997). *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data*. Princeton University Press.

King, G., Rosen, O., and Tanner, M. A. (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research*, 28(1):61–90.

Klein, M., Broadwell, P., Farb, S. E., and Grappone, T. (2018). Comparing published scientific journal articles to their pre-print versions. *International Journal on Digital Libraries*.

Knies, R. (2014). Making Cortana the researcher's dream assistant. https://www.microsoft.com/en-us/research/blog/making-cortana-the-researchers-dream-assistant/. Accessed 2018-04-27.

Knoepfler, P. (2015). Reviewing post-publication peer review. *Trends in Genetics*, 31(5):221–223.

Kousha, K., Thelwall, M., and Abdoli, M. (2018). Can Microsoft Academic assess the early citation impact of in-press articles? a multi-discipline exploratory analysis. *Journal of Informetrics*, 12(1):287–298.

Koya, K. and Chowdhury, G. (2017). Metric-based vs peer-reviewed evaluation of a research output: Lesson learnt from UK's national research assessment exercise. *PLOS ONE*, 12(7):e0179722.

Kuha, J. and Firth, D. (2011). On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models. *Computational Statistics & Data Analysis*, 55(1):375–388.

Lancichinetti, A., Fortunato, S., and Kertész, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Larivière, V., Haustein, S., and Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLOS ONE*, 10(6):e0127502.

Leicht, E. A. and Newman, M. E. J. (2008). Community structure in directed networks. *Physical Review Letters*, 100(11).

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123.

Leskovec, J., Lang, K. J., and Mahoney, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World Wide Web*, pages 631–640. ACM.

Leydesdorff, L. and Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology*, 62(2):217–229.

Leydesdorff, L., Bornmann, L., and Wagner, C. S. (2017). Generating clustered journal maps: an automated system for hierarchical classification. *Scientometrics*, 110(3):1601–1614.

Leydesdorff, L., Bornmann, L., and Zhou, P. (2016). Construction of a pragmatic base line for journal classifications and maps

based on aggregated journal-journal citation relations. *Journal of Informetrics*, 10(4):902–918.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., and de Nooy, W. (2013). Field-normalized impact factors (IFs): A comparison of rescaling and fractionally counted IFs. *Journal of the American Society for Information Science and Technology*, 64(11):2299–2309.

Loeffler, D. (2019). Personal communication.

López-Cózar, E. D., Robinson-García, N., and Torres-Salinas, D. (2013). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, 65(3):446–454.

MacRoberts, M. and MacRoberts, B. (2009). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, 61(1):1–12.

MacRoberts, M. H. and MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36(3):435–444.

MacRoberts, M. H. and MacRoberts, B. R. (2017). The mismeasure of science: citation analysis. *Journal of the Association for Information Science and Technology*, 69(3):474–482.

Malliaros, F. D. and Vazirgiannis, M. (2013). Clustering and community detection in directed networks: a survey. *Physics Reports*, 533(4):95–142.

Mansfield, N. J. (2016). Ranking of design journals based on results of the UK research excellence framework: Using REF as referee. *The Design Journal*, 19(6):903–919.

Marques, M., Powell, J. J., Zapp, M., and Biesta, G. (2017). How does research evaluation impact educational research? Exploring intended and unintended consequences of research assessment in the United Kingdom, 1986–2014. *European Educational Research Journal*, 16(6):820–842.

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., and López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4):1160–1177.

Matthews, D. (2018). French say 'no deal' to Springer as journal fight spreads. `https://www.timeshighereducation.com/fr/news/french-say-no-deal-springer-journal-fight-spreads`. Accessed 2018-05-08.

Maystre, L. and Grossglauser, M. (2015). Fast and accurate inference of Plackett–Luce models. In *Advances in Neural Information Processing Systems*, pages 172–180.

Mazzarol, T. and Soutar, G. N. (2011). Why the ERA had to change and what we should do next. http://theconversation.com/why-the-era-had-to-change-and-what-we-should-do-next-1874. Accessed 2020-02-11.

McCullagh, P. (1982). Some applications of quasisymmetry. *Biometrika*, 69(2):303–308.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Taylor & Francis Ltd, second edition edition.

McDaid, A. F., Murphy, T. B., Friel, N., and Hurley, N. J. (2013). Improved bayesian inference for the stochastic block model with application to large networks. *Computational Statistics & Data Analysis*, 60:12–31.

McVeigh, M. E. and Mann, S. J. (2009). The journal impact factor denominator. *JAMA*, 302(10):1107.

Mirshahvalad, A., Beauchesne, O. H., Archambault, E., and Rosvall, M. (2013). Resampling effects on significance analysis of network clustering and ranking. *PLoS ONE*, 8(1):e53943.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3):265–277.

Mongeon, P. and Paul-Hus, A. (2015). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1):213–228.

Monroe, R. (2008). xkcd: Purity. https://xkcd.com/435/. Accessed 2016-11-26.

Moody, G. (2013). Re-inventing academic publishing: 'diamond' open access titles that are free to read and free to publish. https://www.techdirt.com/articles/20130121/09203321740/re-inventing-academic-publishing-diamond-open-access-titles-that-are-free-to-read-free-to-publish.shtml. Accessed 2018-05-08.

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B. (2015a). Predicting results of the Research Excellence Framework using departmental *h*-index. *Scientometrics*, 102(3):2165–2180.

Mryglod, O., Kenna, R., Holovatch, Y., and Berche, B. (2015b). Predicting results of the Research Excellence Framework using departmental h-index: revisited. *Scientometrics*, 104(3):1013–1017.

Muschelli, J. (2019). *rscopus: Scopus Database 'API' Interface*. R package version 0.6.6.

Negahban, S., Oh, S., and Shah, D. (2012). Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482.

Neuhaus, C. and Daniel, H.-D. (2006). Data sources for performing citation analysis: an overview. *Journal of Documentation*, 64(2):193–210.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press.

Newman, M. E. (2004a). Analysis of weighted networks. *Physical Review E*, 70(5).

Newman, M. E. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6):066133.

Newman, M. E. (2012). Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Neylon, C. and Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biology*, 7(11):e1000242.

Nielsen, A. and Weber, M. (2015). Computing the nearest reversible markov chain. *Numerical Linear Algebra with Applications*, 22(3):483–499.

Normand, S. (2018). Is diamond open access the future of open access? *The iJournal: Graduate Student Journal of the Faculty of Information*, 3(2).

Oswald, A. J. (2007). An examination of the reliability of prestigious scholarly journals: evidence and implications for decision-makers. *Economica*, 74(293):21–31.

Oswald, A. J. (2019). Personal communication.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web. Technical report, Stanford InfoLab.

Palacios-Huerta, I. and Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, 72(3):963–977.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Peng, L. and Carvalho, L. (2016). Bayesian degree-corrected stochastic blockmodels for community detection. *Electronic Journal of Statistics*, 10(2):2746–2779.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312.

PLoS Medicine Editors (2006). The impact factor game. *PLoS Medicine*, 3(6):e291.

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Bachelier, Paris.

Poynder, R. (2014). The subversive proposal at 20. https://poynder.blogspot.co.uk/2014/06/the-subversive-proposal-at-20.html. Accessed 2018-05-04.

Prat-Pérez, A., Dominguez-Sal, D., and Larriba-Pey, J.-L. (2014). High quality, scalable and parallel community detection for large real graphs. In *Proceedings of the 23rd international conference on the World Wide Web*, pages 225–236. ACM.

Prins, A. A., Costas, R., van Leeuwen, T. N., and Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation*, 25(3):264–270.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25(4):348–349.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reed, M. and Kerridge, S. (2017). How much was an impact case study worth in the UK Research Excellence Framework? *Fast Track Impact*, 1:47–49. Accessed 2019-04-09.

REF (2015). REF frequently asked questions. https://www.ref.ac.uk/2014/faq/. Accessed 2020-03-26.

REF (2019). What is the REF? https://www.ref.ac.uk/about/what-is-the-ref/. Accessed 2019-03-07.

Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1).

Rice, C. (2013). Open access publishing hoax: what Science magazine got wrong. https://www.theguardian.com/higher-education-network/blog/2013/oct/04/science-hoax-peer-review-open-access. Accessed 2018-05-04.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rosen, O., Jiang, W., King, G., and Tanner, M. A. (2001). Bayesian and frequentist inference for ecological inference: the R×C case. *Statistica Neerlandica*, 55(2):134–156.

Rosenman, E. (2019). Some new results for poisson binomial models.

Rosenman, E. and Viswanathan, N. (2018). Using Poisson binomial GLMs to reveal voter preferences.

Rossner, M., Epps, H. V., and Hill, E. (2007). Show me the data. *The Journal of Cell Biology*, 179(6):1091–1092.

Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, 178(1):13–23.

Rosvall, M. and Bergstrom, C. T. (2007). An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105:1118–1123.

Rosvall, M. and Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE*, 5(1):e8694.

Schmitt, J. (2014). Academic journals: The most profitable obsolete technology in history. `https://www.huffingtonpost.com/jason-schmitt/academic-journals-the-mos_1_b_6368204.html`. Accessed 2018-04-30.

Scott, P. (2019). Personal communication.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079):497.

Sgroi, D. (2019). Personal communication.

Shafee, T., Das, D., Masukume, G., and Häggström, M. (2017). WikiJournal of Medicine, the first Wikipedia-integrated academic journal. *WikiJournal of Medicine*, 4(1).

Shah, B. K. (1973). On the distribution of the sum of independent integer valued random variables. *The American Statistician*, 27(3):123–124.

Sharp, W. and Markham, T. (2000). Quasi-symmetry and reversible markov sequences in sedimentary sections. *Mathematical Geology*, 32(5):561–579.

Shavell, S. (2010). Should copyright of academic works be abolished? *Journal of Legal Analysis*, 2(1):301–358.

Shema, H. (2013). What's wrong with citation analysis? `https://blogs.scientificamerican.com/information-culture/whats-wrong-with-citation-analysis/`. Accessed 2018-05-01.

Shotton, D. (2013). Publishing: Open citations. *Nature*, 502(7471):295–297.

Silver, A. (2017). Controversial website that lists 'predatory' publishers shuts down. *Nature*.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., and Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*. ACM Press.

Sipido, K. R., Gal, D., Luttun, A., Janssens, S., Sampaolesi, M., and Holvoet, P. (2017). Peer review: (r)evolution needed. *Cardiovascular Research*, 113(13):e54–e56.

Snijders, T. A. and Borgatti, S. P. (1999). Non-parametric standard errors and tests for network statistics. *Connections*, 22(2):161–170.

Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.

Stefaner, M. (2009). The emergence of neuroscience. `http://well-formed-data.net/archives/331/neuroscience-infoporn`.

Stein, D. L. and Newman, C. M. (2012). *Spin Glasses and Complexity*, pages 1–14. University Press Group Ltd.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, 9(1):94–108.

Stockhammer, E., Dammerer, Q., and Kapur, S. (2017). The Research Excellence Framework 2014, journal ratings and the marginalization of heterodox economics. resreport 1715, Post-Keynesian Economics Society.

Straumsheim, C. (2016). How to measure impact. `https://www.insidehighered.com/news/2016/12/14/exploring-citescore-elseviers-new-journal-impact-metrics`. Accessed 2019-08-01.

Szabó, Z., Sriperumbudur, B., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40.

Thelwall, M. (2017a). Does Microsoft Academic find early citations? *Scientometrics*, 114(1):325–334.

Thelwall, M. (2017b). Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals. *Journal of Informetrics*, 11(4):1201–1212.

Thelwall, M. (2018a). Can Microsoft Academic be used for citation analysis of preprint archives? The case of the Social Science Research Network. *Scientometrics*, 115(2):913–928.

Thelwall, M. (2018b). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, 12(1):1–9.

Traag, V. A. and Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications*, 5(1).

van Eck, N. J. and Waltman, L. (2007). VOS: A new method for visualizing similarities between objects. In *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 299–306. Springer Nature.

van Eck, N. J. and Waltman, L. (2009). How to normalize cooccurrence data? an analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, 60(8):1635–1651.

van Eck, N. J., Waltman, L., Dekker, R., and van den Berg, J. (2010a). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, 61(12):2405–2416.

van Eck, N. J., Waltman, L., Noyons, E. C. M., and Buter, R. K. (2010b). Automatic term identification for bibliometric mapping. *Scientometrics*, 82(3):581–596.

Van Noorden, R. (2014). The decline and fall of Microsoft Academic Search. `http://blogs.nature.com/news/2014/05/the-decline-and-fall-of-microsoft-academic-search.html`. Accessed 2018-04-27.

Van Noorden, R. (2016). Controversial impact factor gets a heavyweight rival. *Nature*, 540(7633):325–326.

van Wesel, M. (2015). Evaluation by citation: Trends in publication behavior, evaluation criteria, and the strive for high impact publications. *Science and Engineering Ethics*.

Vanclay, J. K. (2011). An evaluation of the Australian Research Council's journal ranking. *Journal of Informetrics*, 5(2):265–274.

Varin, C., Cattelan, M., and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(1):1–63.

Verma, I. M. (2015). Impact, not impact factor. *Proceedings of the National Academy of Sciences*, 112(26):7875–7876.

Vigna, S. (2016). Spectral ranking. *Network Science*, 4(4):433–445.

Wakefield, J. (2005). Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–445.

Waltman, L. and van Eck, N. J. (2010). The relation between Eigenfactor, audience factor, and influence weight. *Journal of the American Society for Information Science and Technology*, 61(7):1476–1486.

Waltman, L., van Eck, N. J., and Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4(4):629–635.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2):295–312.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, 61(3):439–447.

West, J. D. (2010). *Eigenfactor: ranking and mapping scientific knowledge*. PhD thesis, University of Washington.

Wilhite, A. W. and Fong, E. A. (2012). Coercive citation in academic publishing. *Science*, 335(6068):542–543.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1151–1158, Madison, WI, USA. Omnipress.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., and Johnson, B. (2015). The Metric Tide: Report of the independent review of the role of metrics in research assessment and management. `https://responsiblemetrics.org/the-metric-tide/`. Accessed 2019-04-09.

Xu, R. and Wunsch, D. (2008). *Clustering*. Wiley-IEEE Press.

Yan, Z. (2017). *How does the REF panel perceive journals? A new approach to estimating ordinal response model with censored outcomes*. Phd thesis, Department of Economics, University of Warwick.

A
Appendix

Table A.1: Distribution of Economics and Econometrics REF2014 submissions by containing journal (named titles contained $\geq$ 20 submissions)

Volume title	Outputs	%
American Economic Review	104	4.0
The Economic Journal	103	4.0
Journal of Econometrics	93	3.6
Journal of Economic Theory	81	3.1
Games and Economic Behavior	78	3.0
Econometrica	68	2.6
Journal of the European Economic Association	65	2.5
Review of Economic Studies	63	2.4
Economics Letters	62	2.4
Review of Economics and Statistics	58	2.2
Journal of Public Economics	57	2.2
European Economic Review	51	2.0
Economic Theory	48	1.8
Journal of Development Economics	47	1.8
Journal of Economic Dynamics and Control	44	1.7
Journal of Economic Behavior & Organization	42	1.6
Journal of Monetary Economics	42	1.6
Econometric Theory	35	1.3
Journal of International Economics	35	1.3
Journal of Health Economics	33	1.3
Journal of Money, Credit and Banking	32	1.2
Quarterly Journal of Economics	29	1.1
International Economic Review	28	1.1
Oxford Bulletin of Economics and Statistics	28	1.1
Canadian Journal of Economics	25	1.0
Journal of Applied Econometrics	24	0.9
Oxford Economic Papers	24	0.9
Journal of Banking & Finance	23	0.9
Journal of Political Economy	22	0.8
Conference proceedings	2	0.1
Other journals	944	36.3
Other outputs	210	8.1

Table A.2: Distribution of Physics REF2014 submissions by containing journal (named titles contained ≥ 30 submissions)

Volume title	Outputs	%
Physical Review Letters	1227	19.0
Monthly Notices of the Royal Astronomical Society	678	10.5
The Astrophysical Journal	393	6.1
Physical Review D	281	4.4
Physical Review B	242	3.8
Journal of High Energy Physics	226	3.5
Nature	207	3.2
Astronomy and Astrophysics	196	3.0
Physics Letters B	189	2.9
Science	175	2.7
Applied Physics Letters	123	1.9
Nature Physics	96	1.5
Physical Review A	92	1.4
The European Physical Journal C	87	1.3
Nature Communications	85	1.3
Journal of Geophysical Research	81	1.3
Optics Express	81	1.3
Proceedings of the National Academy of Sciences	81	1.3
New Journal of Physics	76	1.2
Nano Letters	70	1.1
Nature Materials	65	1.0
Advanced Materials	58	0.9
Physical Review C	55	0.9
Journal of the American Chemical Society	52	0.8
Journal of Cosmology and Astroparticle Physics	46	0.7
Nature Photonics	46	0.7
Monthly Notices of the Royal Astronomical Society: Letters	45	0.7
Nuclear Instruments and Methods in Physics Research	45	0.7
Journal of Instrumentation	41	0.6
Nature Nanotechnology	37	0.6
Advanced Functional Materials	35	0.5
ACS Nano	30	0.5
Journal of Physics: Condensed Matter	30	0.5
The Astrophysical Journal Supplement Series	30	0.5
Conference proceedings	18	0.3
Other journals	1075	16.7
Other outputs	52	0.8

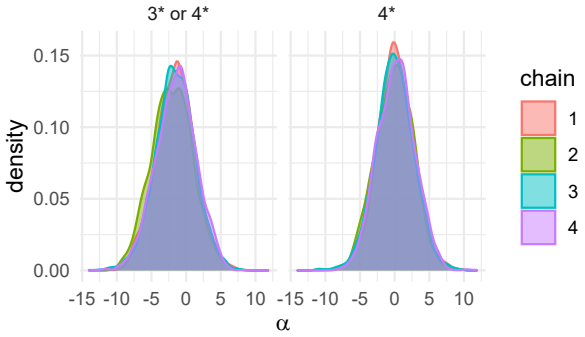
Table A.3: Distribution of Mathematical Sciences REF2014 submissions by containing journal (named titles contained ≥ 30 submissions)

Volume title	Outputs	%
Journal of Fluid Mechanics	254	3.6
Physical Review Letters	209	3.0
Journal of High Energy Physics	159	2.3
Communications in Mathematical Physics	140	2.0
Proceedings of the Royal Society A	126	1.8
Advances in Mathematics	116	1.7
Journal of Physics A	112	1.6
Physical Review E	110	1.6
Physical Review D	107	1.5
Journal of Algebra	83	1.2
Proceedings of the London Mathematical Society	70	1.0
The Annals of Probability	70	1.0
Journal of Functional Analysis	66	0.9
Nonlinearity	66	0.9
Transactions of the American Mathematical Society	66	0.9
The Annals of Applied Probability	61	0.9
Biometrika	57	0.8
SIAM Journal on Numerical Analysis	55	0.8
Journal of the London Mathematical Society	54	0.8
International Mathematics Research Notices	52	0.7
Mathematische Annalen	50	0.7
Archive for Rational Mechanics and Analysis	49	0.7
JRSS Series B (Statistical Methodology)	48	0.7
Crelles Journal	47	0.7
JRSS Series C (Applied Statistics)	47	0.7
Physics of Fluids	45	0.6
Journal of Mathematical Physics	44	0.6
Annals of Mathematics	42	0.6
SIAM Journal on Applied Mathematics	42	0.6
Probability Theory and Related Fields	41	0.6
SIAM Journal on Scientific Computing	41	0.6
Stochastic Processes and their Applications	41	0.6
Journal of Differential Equations	40	0.6
Proceedings of the National Academy of Sciences	40	0.6
Bulletin of the London Mathematical Society	39	0.6
Duke Mathematical Journal	38	0.5
Journal of Pure and Applied Algebra	38	0.5
SIAM Journal on Mathematical Analysis	38	0.5
Geometric and Functional Analysis	37	0.5
Journal of Mathematical Biology	37	0.5
Journal of the American Statistical Association	37	0.5
The Astrophysical Journal	37	0.5
Bulletin of Mathematical Biology	36	0.5

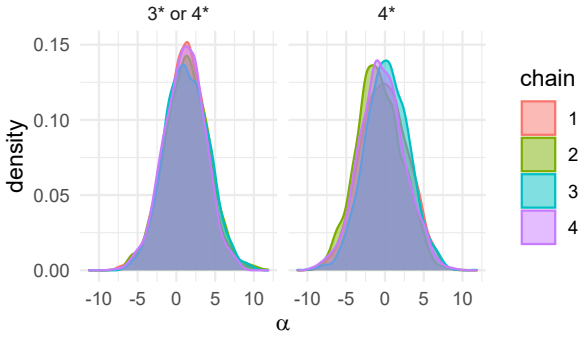
Inventiones mathematicae	36	0.5
The Annals of Statistics	35	0.5
Mathematische Zeitschrift	34	0.5
Physica D	34	0.5
European Journal of Operational Research	33	0.5
Nuclear Physics B	33	0.5
Biometrics	32	0.5
Journal of Computational Physics	32	0.5
Journal of Theoretical Biology	32	0.5
Compositio Mathematica	31	0.4
Journal of Statistical Physics	31	0.4
Electronic Journal of Probability	30	0.4
Geometry & Topology	30	0.4
The Annals of Applied Statistics	30	0.4
Conference proceedings	17	0.2
Other journals	3291	47.1
Other outputs	246	3.5

Table A.4: Distribution of Chemistry REF2014 submissions by containing journal (named titles contained ≥ 30 submissions)

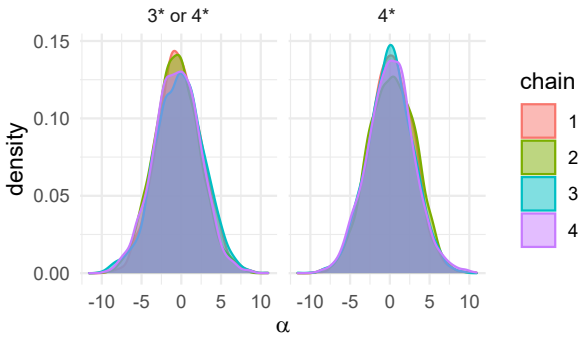
Volume title	Outputs	%
Journal of the American Chemical Society	690	14.7
Angewandte Chemie International Edition	472	10.0
Chemical Communications	258	5.5
Proceedings of the National Academy of Sciences	142	3.0
Chemistry - A European Journal	138	2.9
Physical Chemistry Chemical Physics	126	2.7
Nature Chemistry	119	2.5
The Journal of Chemical Physics	116	2.5
Physical Review Letters	112	2.4
Chemical Science	94	2.0
Science	90	1.9
The Journal of Physical Chemistry C	80	1.7
Dalton Transactions	76	1.6
Inorganic Chemistry	71	1.5
The Journal of Organic Chemistry	70	1.5
Organic Letters	68	1.4
Chemistry of Materials	57	1.2
Organic & Biomolecular Chemistry	55	1.2
Advanced Materials	54	1.1
Analytical Chemistry	53	1.1
Nature	53	1.1
Langmuir	51	1.1
Journal of Materials Chemistry	50	1.1
ACS Nano	49	1.0
The Journal of Physical Chemistry A	49	1.0
Nature Materials	47	1.0
Atmospheric Chemistry and Physics	44	0.9
Soft Matter	43	0.9
The Journal of Physical Chemistry B	43	0.9
Organometallics	41	0.9
Physical Review B	41	0.9
Advanced Functional Materials	38	0.8
Journal of Medicinal Chemistry	37	0.8
Journal of Biological Chemistry	34	0.7
Nano Letters	32	0.7
Nature Communications	31	0.7
Conference proceedings	2	0.0
Other journals	1064	22.6
Other outputs	8	0.2



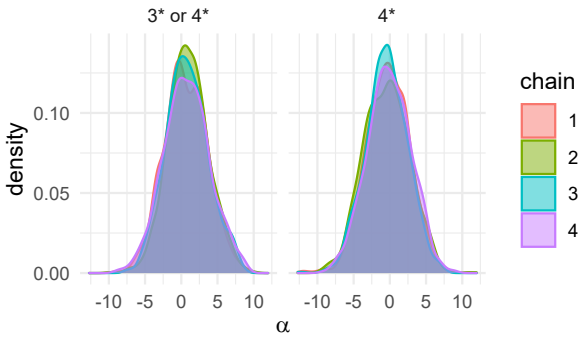
(a) Economics and Econometrics



(b) Mathematical Sciences

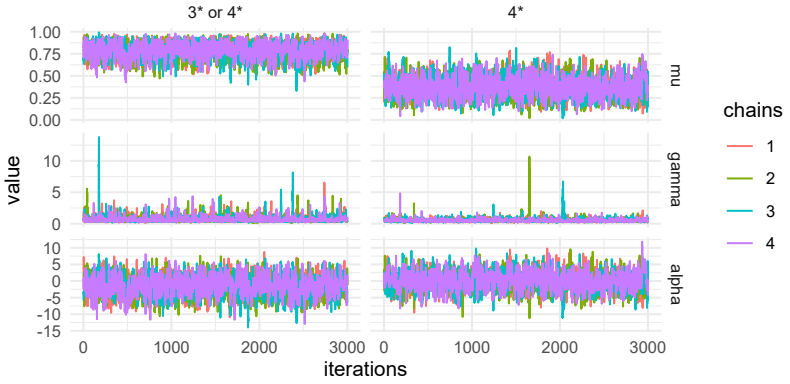


(c) Physics

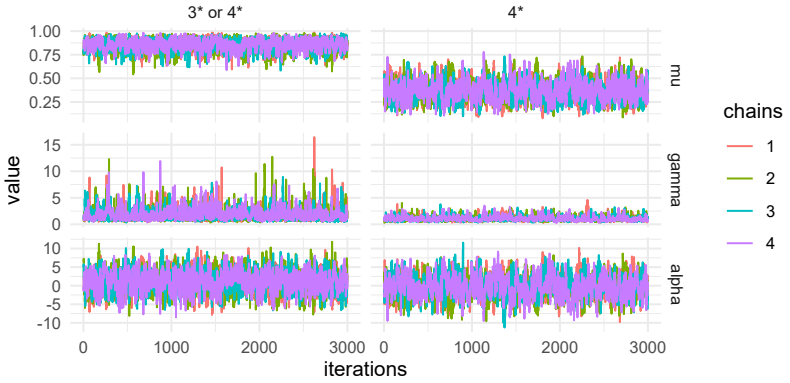


(d) Chemistry

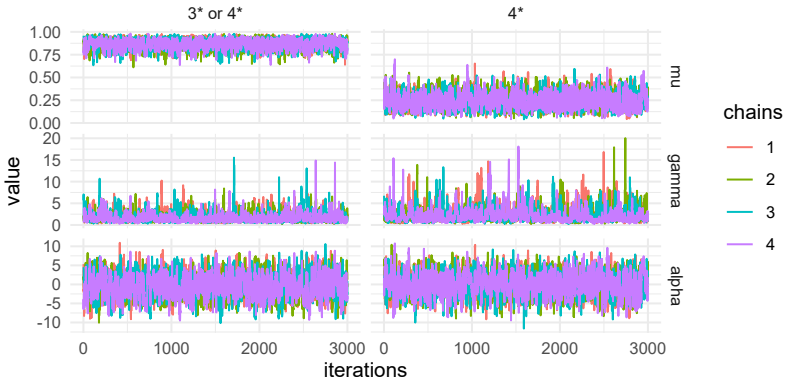
Figure A.1: Marginal density of α hyper-parameter for four chains of Hamiltonian Monte Carlo, run on 4* and 3*+ profiles for each field. The prior for α is a normal distribution with mean zero and standard deviation 3



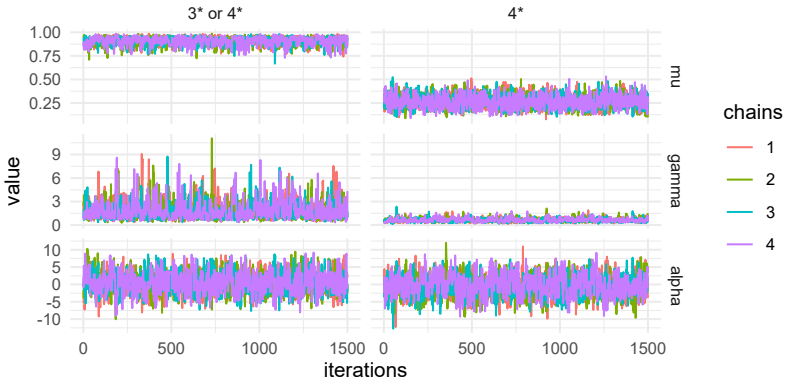
(a) Economics and Econometrics



(b) Mathematical Sciences



(c) Physics



(d) Chemistry

Figure A.2: Hamiltonian Monte Carlo trace plots for different parameters in the Poisson binomial model, run on 4* and 3*+ profiles for each field